

10-01-04

JFW

PTO/SB/21 (04-04)

**TRANSMITTAL
FORM**

(to be used for all correspondence after initial filing)

TRANSMITTAL FORM (to be used for all correspondence after initial filing)	Application Number	10/816,572	
	Filing Date	March 31, 2004	
	First Named Inventor	IWAMURA, Takashige	
	Art Unit	2181	
	Examiner Name	Unassigned	
Total Number of Pages in This Submission	12	Attorney Docket Number	16869P-112100US

ENCLOSURES (Check all that apply)

<input checked="" type="checkbox"/> Fee Transmittal Form (in duplicate) <input type="checkbox"/> Fee Attached <input type="checkbox"/> Amendment/Reply <input type="checkbox"/> After Final <input type="checkbox"/> Affidavits/declaration(s) <input type="checkbox"/> Extension of Time Request <input type="checkbox"/> Express Abandonment Request <input type="checkbox"/> Information Disclosure Statement <input type="checkbox"/> Certified Copy of Priority Document(s) <input type="checkbox"/> Response to Missing Parts/Incomplete Application <input type="checkbox"/> Response to Missing Parts under 37 CFR 1.52 or 1.53	<input type="checkbox"/> Drawing(s) <input type="checkbox"/> Licensing-related Papers <input checked="" type="checkbox"/> Petition To Make Special (9 pages) <input type="checkbox"/> Petition to Convert to a Provisional Application <input type="checkbox"/> Power of Attorney, Revocation Change of Correspondence Address <input type="checkbox"/> Terminal Disclaimer <input type="checkbox"/> Request for Refund <input type="checkbox"/> CD, Number of CD(s) _____	<input type="checkbox"/> After Allowance Communication to Technology Center (TC) <input type="checkbox"/> Appeal Communication to Board of Appeals and Interferences <input type="checkbox"/> Appeal Communication to TC (Appeal Notice, Brief, Reply Brief) <input type="checkbox"/> Proprietary Information <input type="checkbox"/> Status Letter <input checked="" type="checkbox"/> Other Enclosure(s) (please identify below): Return Postcard Nine (9) cited references
Remarks The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.		

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual name	Townsend and Townsend and Crew LLP Chun-Pok Leung	Reg. No. 41,405
Signature		
Date	October 1, 2004	

CERTIFICATE OF TRANSMISSION/MAILING

Express Mail Label: EV 530887089 US			
I hereby certify that this correspondence is being deposited with the United States Postal Service with "Express Mail Post Office to Address" service under 37 CFR 1.10 on this date October 1, 2004 and is addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.			
Typed or printed name	Joy Salvador		
Signature		Date	October 1, 2004

FEE TRANSMITTAL for FY 2004

Effective 10/01/2003. Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT (\$) 130.00

Complete if Known

Application Number	10/816,572
Filing Date	March 31, 2004
First Named Inventor	IWAMURA, Takashige
Examiner Name	Unassigned
Art Unit	2181
Attorney Docket No.	16869P-112100US

METHOD OF PAYMENT (check all that apply)

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None

☒ Deposit Account:
Deposit
Account
Number

20-1430

Deposit
Account
Name

Townsend and Townsend and Crew LLP

The Director is authorized to: (check all that apply)

☒ Charge fee(s) indicated below ☒ Credit any overpayments

☒ Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, except for the filing fee to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity		Small Entity		Fee Description	Fee Paid
Fee Code	Fee (\$)	Fee Code	Fee (\$)		
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	

SUBTOTAL (1)

(\$0.00)

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

Total Claims		Extra Claims		Fee from below		Fee Paid
Independent Claims		** =		X		
Multiple Dependent		** =		X		

Large Entity		Small Entity		Fee Description
Fee Code	Fee (\$)	Fee Code	Fee (\$)	
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2)

(\$0.00)

**or number previously paid, if greater; For Reissues, see above

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity		Small Entity		Fee Description	Fee Paid
Fee Code	Fee (\$)	Fee Code	Fee (\$)		
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive - unavoidable	
1453	1,330	2453	665	Petition to revive - unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	130
1807	50	1807	50	Petitions related to provisional applications	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR § 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37 CFR § 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid

SUBTOTAL (3)

(\$130.00)

SUBMITTED BY

Name (Print/Type)

Chun-Pok Leung

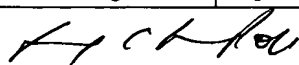
Registration No. (Attorney/Agent)

41,405

Telephone

650-326-2400

Signature



Date

October 1, 2004

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.



PATENT
Attorney Docket No.: 16869P-112100US
Client Ref. No.: 340301728US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

TAKASHIGE IWAMURA *et al.*

Application No.: 10/816,572

Filed: March 31, 2004

For: REMOTE COPY NETWORK

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2181

Confirmation No.: 1244

**PETITION TO MAKE SPECIAL FOR
NEW APPLICATION UNDER M.P.E.P.
§ 708.02, VIII & 37 C.F.R. § 1.102(d)**

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of \$130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

10/06/2004 SSITHIB1 00000093 201430 10816572

01 FC:1460 130.00 DA

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicants will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search and a computer database search. The searches were performed on or around September 2, 2004, and were conducted by a professional search firm, Kramer & Amado, P.C. The classification search covered Classes 711 (subclasses 11, 112, 161, and 162), 714 (subclasses 6 and 13), and 709 (subclasses 203, 217, and 219) for the U.S. and foreign subclasses identified above. The computer database search was conducted on the USPTO systems EAST and WEST. The inventors further provided seven references considered most closely related to the subject matter of the present application (see references #3-9 below), which were cited in the Information Disclosure Statement filed with the application on March 31, 2004.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

- (1) U.S. Patent Publication No. 2004/0044865 A1;
- (2) U.S. Patent Publication No. 2004/0039888 A1;
- (3) U.S. Patent Publication No. 2003/0051111 A1;
- (4) U.S. Patent No. 6,209,002 B1;
- (5) U.S. Patent No. 5,734,818;
- (6) Japanese Patent Publication No. JP 2003-122509;
- (7) Japanese Patent Publication No. JP 2000-305856;
- (8) Japanese Patent Publication No. JP 07-244597; and
- (9) European Patent Publication No. EP 1049016 A2.

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to an information processing system comprising storage device, and more specifically, to remote copying and disaster recovery technology, executed by a remote copy network comprising two or more storage devices and two or more network devices.

Independent claim 1 recites a remote copy network system having a first storage system and a second storage system, including a first edge device coupled to the first storage system, a second edge device coupled to the second storage system, the first edge device and the second edge device being coupled by a network. The first edge device receives a remote copy I/O request to copy data to the second storage system from the first storage system, and sends a response to the received remote copy I/O request to the first storage system. After sending a response to the remote copy I/O request, the first edge device sends to the second edge device log information having the remote copy I/O request and a sequential number indicating the order of reception of the remote copy I/O request. The second edge device extracts the remote copy I/O request from the received log information, and sends the extracted remote copy I/O request to the second storage system according to the order indicated by the sequential number in the log information.

Independent claim 11 recites a relay device, coupled to a first storage system which relays a remote copy of data from the first storage system to a second storage system. The relay device comprises a first interface coupled to the first storage system; a second interface coupled to another relay device via a network, wherein the other relay device is coupled to the second storage system; a processor; and a memory. The first interface receives a remote copy I/O request for remote copying data from the first storage system to the second storage system, and returns a response to the remote copy I/O request to the first storage system. The processor creates and stores in the memory log information having a remote copy I/O request and a sequential number indicating the order of reception of the remote copy I/O request. The second interface portion, after returning a response to the remote copy I/O request, sends the created log information to the other relay device. The second interface

portion receives a response to the log information. The processor deletes from the memory the log information corresponding to the received response.

Independent claim 17 recites a relay device, coupled to a second storage system, for relaying remote copy data from a first storage system to the second storage system. The relay device comprises a first interface portion, coupled to another relay device via a network, wherein the other relay device is coupled to the first storage system; a second interface, coupled to the second storage system; and a processor. The first interface portion receives, from the other relay device, log information having a remote copy I/O request for remote copying data from the first storage system to the second storage system and a sequential number indicating the order of reception at the other relay device of the remote copy I/O request. The processor acquires the remote copy I/O request from the received log information. The second interface portion sends the acquired remote copy I/O request, in the order of the sequential number comprised in the log information, to the second storage system.

One benefit that may be derived is that device ownership costs and management costs when executing multi-hop remote copying between storage devices can be reduced. See specification at page 3, lines 19-21.

B. Discussion of the References

None of the following references disclose that, after sending a response to the remote copy I/O request, the first edge device sends to the second edge device log information having the remote copy I/O request and a sequential number indicating the order of reception of the remote copy I/O request; and that the second edge device extracts the remote copy I/O request from the received log information, and sends the extracted remote copy I/O request to the second storage system according to the order indicated by the sequential number in the log information.

The references further fail to teach a relay device coupled to a first storage system which relays a remote copy of data from the first storage system to a second storage system, wherein the processor creates and stores in the memory log information having a remote copy I/O request and a sequential number indicating the order of reception of the

remote copy I/O request; the second interface portion, after returning a response to the remote copy I/O request, sends the created log information to the other relay device; the second interface portion receives a response to the log information; and the processor deletes from the memory the log information corresponding to the received response.

In addition, the references do not disclose a relay device coupled to a second storage system for relaying remote copy data from a first storage system to the second storage system, wherein the first interface portion receives, from the other relay device, log information having a remote copy I/O request for remote copying data from the first storage system to the second storage system and a sequential number indicating the order of reception at the other relay device of the remote copy I/O request; the processor acquires the remote copy I/O request from the received log information; and the second interface portion sends the acquired remote copy I/O request, in the order of the sequential number comprised in the log information, to the second storage system.

1. U.S. Patent Publication No. 2004/0044865 A1

This reference discloses a method for transaction command ordering in a remote data replication system. A disaster-tolerant data backup and remote copy system which is implemented as a controller-based replication of one or more LUNs (logical units) between two remotely separated pairs of array controllers connected by redundant links.

2. U.S. Patent Publication No. 2004/0039888 A1

This reference discloses an automated storage replication processing when a disaster occurs, the state of the replication processing is determined, and a restart copy of the data is made available from the recover site. Processing continues based on whether protection mode is desired such that the system executes using the recovery site as the restart with a replicated copy of the data. A multi-hop configuration may use consistency group technology to provide protection of the production site in an offsite secondary site or "bunker site." See [0098].

3. U.S. Patent Publication No. 2003/0051111 A1

This reference discloses a remote copy control method, a storage sub-system with the method, and a large area data storage system using them. With the multi-hop method either synchronous transfers or asynchronous transfers are arbitrarily set for communication among the storage sub-systems. See [0198].

4. U.S. Patent No. 6,209,002 B1

This reference discloses a data storage facility for transferring data from a data altering apparatus, such as a production data processing site to a remote data receiving site. The data storage facility includes a first data store for recording each change in the data generated by the data altering apparatus. A register set records each change on a track-by-track basis. A second data store has first and second operating modes. During a first operating mode the second data store becomes a mirror of the first data store. During a second operating mode the second data store ceases to act as a mirror and becomes a source for a transfer of data to the data receiving site. Only information that has been altered, i.e., specific tracks that have been altered, are transferred during successive operations in the second operating mode. Commands from the local production site initiate the transfers between the first and second operating modes.

5. U.S. Patent No. 5,734,818

This reference relates to a remote data shadowing system that provides storage based, real time disaster recovery capability. Record updates at a primary site cause write I/O operations in a storage subsystem therein. The write I/O operations are time stamped and the time, sequence, and physical locations of the record updates are collected in a primary data mover. The primary data mover groups sets of the record updates and associated control information based upon a predetermined time interval, the primary data mover appending a prefix header to the record (updates thereby forming self describing record sets. The self describing record sets are transmitted to a remote secondary site wherein consistency groups are formed such that the record updates are ordered so that the record updates can be shadowed in an order consistent with the order the record updates cause write I/O operations at the primary site.

6. Japanese Patent Publication No. JP 2003-122509

This reference discloses a remote control method that always hold the sequence of updating data between three or more data centers. Two data centers existing neighboring place are connected using a copy function by simultaneous transfer. One of the data centers and a third data center existing in a remote place are connected by an asynchronous remote copy function, whereby a storage subsystem existing in a neighboring place always ensures the sequence of the data received from a host and the third data center holds the data. Further, each storage subsystem is provided with a function of grasping the progress state of transferring, receiving, and updating the data between the storage subsystems installed in two data centers which do not directly transfer data in normal operation.

Because the method assumes that the owner of the storage devices manages intermediate devices, increases in the costs of device ownership and in management costs are problematic. See present specification at page 3, lines 8.

7. Japanese Patent Publication No. JP 2000-305856

This reference relates to a technique to guarantee the sequence of update and the consistency of data by doubling data between disk subsystems on a main-center side and a remote-center side through gateway subsystems. Data which are written from a host computer 1 are doubled between disk subsystems 3-1, 3-2, . . . , 3-n, and a gateway subsystem 5 and held macroscopically in the same state. The gateway subsystem 5 adds information for holding the sequence of update. Further, the data are doubled between the gateway subsystem 5 and a gateway subsystem 7 by asynchronous remote copying while the sequence of update is guaranteed. The disk subsystems 9-1, 9-2, . . . , 9-n have the data updated in synchronism with the update of the gateway subsystem 7. Those are all actualized only by the function of the disk subsystems and no new software need be introduced.

Because the method assumes that the owner of the storage devices manages intermediate devices, increases in the costs of device ownership and in management costs are problematic. See present specification at page 3, lines 8.

8. Japanese Patent Publication No. JP 07-244597

This reference relates to a technique to provide a remote data shadowing system which provides a real-time disaster recovery function on a storage area base. A write input-output operation is performed in a storage subsystem on the primary side 14 by record update on the primary side 14. A time-stamp is attached to this write input-output operation and the time, order and physical position of the record update are collected in a primary data mover. The primary data mover divides plural sets of record update and their related control information into groups based on prescribed time intervals, adds a prefix header to the record update and thereby forms a self-description record set. The self-description record set is sent to a remote secondary side 15, and such a consistency group is formed that the record update is ordered to be able to shadow the record update in the sequence that matches the sequence where the write input-output operation was performed on the primary side 14 by the record update.

9. European Patent Publication No. EP 1049016 A2

This reference discloses an asynchronous remote copy system that can ensure the data renewal order and data integrity of the disk subsystems and are easy to be incorporated and free from degradation of the process performance by host computers (1, 11). To this end, in the remote copy system for data mirroring, a main center (12) has one gateway subsystem (5) and a remote center (13) has one gateway subsystem (7), and disk subsystems (3, 9) in each center to be remotely copied are connected to the corresponding gateway subsystem. Data is mirrored through synchronous type remote copy between a volume of the disk subsystem of each center to be remotely copied and a desired volume of the corresponding gateway subsystem, and the gateway subsystem of the main center sends the renewal data to the gateway subsystem of the remote center in accordance with the order of renewal of volumes of the gateway subsystem of the main center, to make the gateway subsystem of the remote center reflect the renewal data upon the volumes thereof through asynchronous type remote copy.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,



Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60314038 v1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-122509

(43)Date of publication of application : 25.04.2003

(51)Int.Cl.

G06F 3/06
G06F 12/00

(21)Application number : 2002-019971

(71)Applicant : HITACHI LTD

(22)Date of filing : 29.01.2002

(72)Inventor : NAKANO TOSHIO
NAKAMURA KATSUNORI
OGATA MIKITO
OKAMI YOSHINORI
HIGAKI SEIICHI
ABEI MASARU
KIJIRO SHIGERU

(30)Priority

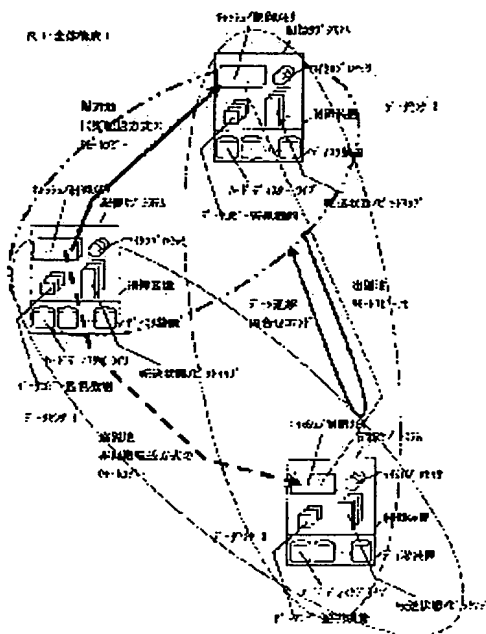
Priority number : 2001240072 Priority date : 08.08.2001 Priority country : JP

(54) REMOTE COPY CONTROL METHOD, STORAGE SUB-SYSTEM USING IT, AND WIDE AREA DATA STORAGE SYSTEM USING THEM

(57)Abstract:

PROBLEM TO BE SOLVED: To always hold the sequence of updating data between three or more data centers.

SOLUTION: Two data centers existing neighboring places are connected using a copy function by simultaneous transfer. One of the data centers and a third data center existing in a remote place are connected by an asynchronous remote copy function, whereby a storage sub-system existing in a neighboring place always ensures the sequence of the data received from a host and the third data center holds the data. Further, each storage sub-system is provided with a function of grasping the progress state of transferring, receiving and updating the data between the storage sub-systems installed in two data centers which do not directly transfer data in normal operation.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開2003-122509

(P2003-122509A)

(43)公開日 平成15年4月25日(2003.4.25)

(51)Int.Cl. ⁷	識別記号	F I	テマコード*(参考)
G 0 6 F 3/06	3 0 4	G 0 6 F 3/06	3 0 4 F 5 B 0 6 5
12/00	5 3 1	12/00	5 3 1 D 5 B 0 8 2
	5 3 3		5 3 3 J

審査請求 未請求 請求項の数34 O L (全 34 頁)

(21)出願番号 特願2002-19971(P2002-19971)
 (22)出願日 平成14年1月29日(2002.1.29)
 (31)優先権主張番号 特願2001-240072(P2001-240072)
 (32)優先日 平成13年8月8日(2001.8.8)
 (33)優先権主張国 日本(JP)

(71)出願人 000005108
 株式会社日立製作所
 東京都千代田区神田駿河台四丁目6番地
 (72)発明者 中野 俊夫
 神奈川県小田原市中里322番地2号 株式
 会社日立製作所RAIDシステム事業部内
 (72)発明者 中村 勝憲
 神奈川県小田原市中里322番地2号 株式
 会社日立製作所RAIDシステム事業部内
 (74)代理人 100071283
 弁理士 一色 健輔

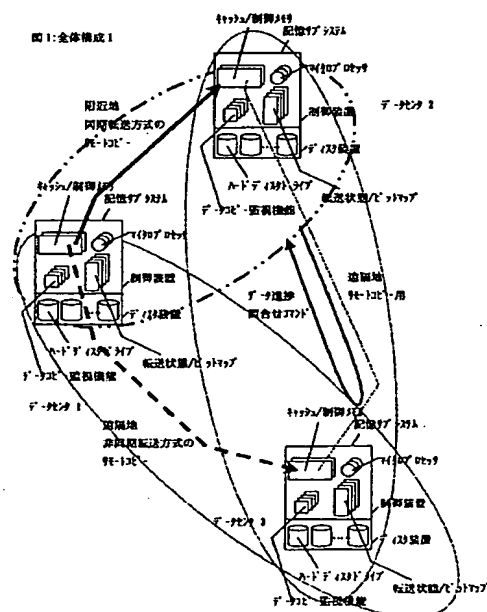
最終頁に続く

(54)【発明の名称】 リモートコピー制御方法、これを用いた記憶サブシステム、及び、これらを用いた広域データス

(57)【要約】 ト修正消システム

【課題】 3以上のデータセンタ間で、常時、データ更新の順序性を保持する。

【解決手段】 付近地に存在する2つのデータセンタ間では同期転送によるコピー機能を用いた接続構成とする。これらのうち1つのデータセンタと、遠隔地に存在する第3のデータセンタとの間には、非同期リモートコピー機能で連結し、付近地に存在する記憶サブシステムがホストから受領したデータの順序性を常時、保証しつつ第3のデータセンタが、そのデータを保持する。更に、正常運用の際には直接にデータ転送を行なわない2つのデータセンタに設置された記憶サブシステムの間で、データ転送・受領・更新の進捗状態を把握する機能を、各記憶サブシステムに設ける。



【特許請求の範囲】

【請求項1】 制御情報を格納する制御メモリと、データを一時的に格納するキャッシュメモリと、これらを制御するマイクロプロセッサにより、前記データを同期転送するN個の転送先と、前記データを非同期転送するM個の転送先とを有する記憶サブシステムのリモートコピー制御方法において、前記制御メモリが、N+M個のデータの転送先に対応する転送状態／ビットマップを格納する第1のステップと、前記転送状態／ビットマップのうち1つに対応する転送状態／ビットマップを有する別の記憶サブシステムであって、直接にデータ転送を行っていないものに対して、データ更新の進捗状態を問合せコマンドを発する第2のステップと、第2のステップの応答を受けて、転送状態／ビットマップを更新する第3のステップとを有することを特徴とするリモートコピー制御方法。

【請求項2】 請求項1記載のリモートコピー制御方法において、第3のステップにおける転送状態／ビットマップの更新は、前記データ更新の進捗状態を問合せコマンドを含む、データブロックの一部に対するデータの更新回数を計数するためのカウンタ値を更新することを含むリモートコピー制御方法。

【請求項3】 制御情報を格納する制御メモリと、データを一時的に格納するキャッシュメモリと、これらを制御するマイクロプロセッサにより、前記データを同期転送する転送先と、前記データを非同期転送する転送先とを有する記憶サブシステムにおいて、直接データの転送を行わない別の記憶サブシステムに対し、データ更新の進捗状態を問合せ機能を有することを特徴とする記憶サブシステム。

【請求項4】 請求項3記載の記憶サブシステムは、更に、自己と、直接データの転送を行わない別の記憶サブシステムにおいて保持する、1組の転送状態／ビットマップを有し、前記進捗状態を問合せ機能により、当該1組の転送状態／ビットマップを更新する記憶サブシステム。

【請求項5】 請求項4記載の記憶サブシステムにおいて、前記1組の転送状態／ビットマップは、データブロックの一部に対するデータの更新回数を計数するためのカウンタ値を格納する領域を有する記憶サブシステム。

【請求項6】 第1のデータセンタに設置された第1の記憶サブシステム、第2のデータセンタに設置された第2の記憶サブシステム、及び、第3のデータセンタに設置された第3の記憶サブシステムを有する広域データストレージシステムにおいて、第1の記憶サブシステムと第2の記憶サブシステムは、

同期転送によりデータが転送され、

第1の記憶サブシステムと第3の記憶サブシステムは、非同期転送によりデータが転送され、第3の記憶サブシステムにおいて、当該転送されたデータの順序性が、常時、保たれることを特徴とする広域データストレージシステム。

【請求項7】 請求項6記載の広域データストレージシステムにおいて、

第2の記憶サブシステムから、同期転送によりデータの転送を受けた第1の記憶サブシステムが、第3の記憶サブシステムに対し、非同期転送により当該データを転送する広域データストレージシステム。

【請求項8】 請求項6記載の広域データストレージシステムにおいて、

第1の記憶サブシステムは、ホストから受領したデータを、同期転送により第2の記憶サブシステムへ転送すると共に、非同期転送により第3の記憶サブシステムにも転送する広域データストレージシステム。

【請求項9】 請求項6記載の広域データストレージシステムにおいて、

第1のデータセンタと第2のデータセンタは附近地に存在し、第1のデータセンタと第3のデータセンタは遠隔地に存在する広域データストレージシステム。

【請求項10】 第1のデータセンタに設置された第1の記憶サブシステム、第2のデータセンタに設置された第2の記憶サブシステム、及び、第3のデータセンタに設置された第3の記憶サブシステムを有する広域データストレージシステムにおける、3つ以上のデータセンタの間のリモートコピー制御方法であって、ホストからのデータを第1の記憶サブシステムが受領する第1のステップと、

第1の記憶サブシステムが、前記ホストからのデータを、第2の記憶サブシステムへ、同期転送する第2のステップと、

第1の記憶サブシステムが、前記ホストからのデータを、第3の記憶サブシステムへ、非同期転送する第3のステップと、

第2の記憶サブシステムから第3の記憶サブシステムへ、前記ホストからのデータが第3の記憶サブシステムへ到着したか否かを問合せ第4のステップとを有するリモートコピー制御方法。

【請求項11】 請求項10記載のリモートコピー制御方法において、更に、

第1のデータセンタが機能停止する第5のステップと、第2の記憶サブシステムから第3の記憶サブシステムへ、第2の記憶サブシステムが保持するデータの一部を転送する第6のステップとを有するリモートコピー制御方法。

【請求項12】 第1のデータセンタに設置された第1の記憶サブシステム、第2のデータセンタに設置された

第2の記憶サブシステム、及び、第3のデータセンタに設置された第3の記憶サブシステムを有する広域データストレージシステムにおける、3つ以上のデータセンタの間のリモートコピー制御方法であって、

ホストからのデータを第1の記憶サブシステムが受領する第1のステップと、

第1の記憶サブシステムが、前記ホストからのデータを、第2の記憶サブシステムへ、同期転送する第2のステップと、

第2の記憶サブシステムが、同期転送された前記ホストからのデータを、第3の記憶サブシステムへ、非同期転送する第3のステップと、

第1の記憶サブシステムから第3の記憶サブシステムへ、前記ホストからのデータが第3の記憶サブシステムへ到着したか否かを問合せる第4のステップとを有するリモートコピー制御方法。

【請求項13】 請求項12記載のリモートコピー制御方法において、更に、第2のデータセンタが機能停止する第5のステップと、

第1の記憶サブシステムから第3の記憶サブシステムへ、第1の記憶サブシステムが保持するデータの一部を転送する第6のステップとを有するリモートコピー制御方法。

【請求項14】 記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおけるリモートコピー制御方法において、

記憶サブシステムが、前記記憶資源にデータを書き込むステップと、

記憶サブシステムが、前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶するステップと、

記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先に送信するステップと、

記憶サブシステムが、前記転送先から送られてくる前記シーケンス番号を受信するステップと、

記憶サブシステムが、記憶している前記対応づけと前記転送先から受信した前記シーケンス番号とに基づいて、前記転送先において未反映となっている書き込みデータを把握するステップと、

を有することを特徴とするリモートコピー制御方法。

【請求項15】 第1の記憶資源に対するデータ書き込み手段を備え第1のサイトに設置された第1の記憶サブシステムと、第1の記憶サブシステムに接続するホストと、第2の記憶資源に対するデータ書き込み手段を備え第2のサイトに設置された第2の記憶サブシステムと、第3の記憶資源に対するデータ書き込み手段を備え第3のサイトに設置された第3の記憶サブシステムと、を有

する広域データストレージシステムにおけるリモートコピー制御方法において、

第1の記憶サブシステムが、前記ホストからの指示により第1の記憶資源に対してデータの書き込みを行うステップと、

第1の記憶サブシステムが、第1の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶するステップと、

第1の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを第2の記憶サブシステムに送信するステップと、

第2の記憶サブシステムが、前記データと前記対応づけとを受信して前記データを第2の記憶資源に記憶し、前記データと前記対応づけとを第3の記憶サブシステムに送信するステップと、

第3の記憶サブシステムが、前記データと前記対応づけとを受信して前記データを第3の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を第1の記憶サブシステムに送信するステップと、

第1の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、第3の記憶資源に未反映となっているデータを把握するステップと、

を有することを特徴とするリモートコピー制御方法。

【請求項16】 前記第2の記憶サブシステムが障害等により使用できなくなった場合に、

前記第1の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第3の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第3の記憶資源に記憶して前記第1の記憶資源と前記第3の記憶資源の内容を同期させるステップと、

第1の記憶サブシステムが、前記ホストからの指示により前記第1の記憶資源に対してデータ書き込みが行われた場合に、前記データ書き込みにより書き込まれたデータと前記第1の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記データと前記書き込み位置情報とを受信して前記データを前記第3の記憶資源の前記書き込み位置情報で特定される位置に記憶するステップと、

を有することを特徴とする請求項15に記載のリモートコピー制御方法。

【請求項17】 前記第2の記憶サブシステムが回復し

再び使用できるようになった場合に、

前記第1の記憶サブシステムが、前記第1の記憶資源に格納されている全データを前記第2の記憶資源に転送するステップと、

第1の記憶サブシステムが、前記ホストからの指示により第1の記憶資源にデータ書き込みを行い、書き込んだデータとシーケンス番号とを第2の記憶サブシステムおよび第3の記憶サブシステムに送信するステップと、

第2の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを第2の記憶資源に記憶するステップと、

前記第2の記憶サブシステムが、前記第2の記憶資源に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶するステップと（位置情報管理テーブル作成）、

前記第3の記憶サブシステムが、記憶サブシステム1から送られてくる前記データと前記シーケンス番号とを受信して前記データを前記第3の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を前記第2の記憶サブシステムに送信するステップと、

前記第1の記憶サブシステムから前記第3の記憶サブシステムへのデータの転送を停止するステップと、

前記第2の記憶サブシステムが、記憶サブシステム3から送られてくる前記シーケンス番号を受信して、前記シーケンス番号と自身が記憶している前記シーケンス番号とこれに対応する書き込み位置情報とに基づいて、前記第3の記憶資源に未反映となっているデータを把握するステップと、

前記第2の記憶サブシステムが、前記把握した前記第3の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第3の記憶資源に記憶して前記第1の記憶資源と前記第3の記憶資源の内容を同期させるステップと、

を有することを特徴とする請求項16に記載のリモートコピー制御方法。

【請求項18】 記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおけるリモートコピー制御方法において、

記憶サブシステムが、前記記憶資源にデータを書き込むステップと、

記憶サブシステムが、前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを生成するステップと、

記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先の一つである転送先Aに送信するステップと、

転送先Aが記憶サブシステムから送信されてくる前記対応づけを受信して記憶するステップと、

転送先Aが別の前記転送先である転送先Bから送られてくる前記シーケンス番号を受信するステップと、

転送先Aが、記憶している前記対応づけと転送先Bから受信した前記シーケンス番号とに基づいて、転送先Bにおいて未反映となっている書き込みデータを把握するステップと、

を有することを特徴とするリモートコピー制御方法。

【請求項19】 第1の記憶資源に対するデータ書き込み手段を備え第1のサイトに設置された第1の記憶サブシステムと、第2の記憶資源に対するデータ書き込み手段を備え第2のサイトに設置された第2の記憶サブシステムと、第2の記憶サブシステムに接続するホストと、第3の記憶資源に対するデータ書き込み手段を備え第3のサイトに設置された第3の記憶サブシステムと、を有する広域データストレージシステムにおけるリモートコピー制御方法であって、

第2の記憶サブシステムが、前記ホストからの指示により第2の記憶資源に対してデータの書き込みを行うステップと、

第2の記憶サブシステムが、前記書き込んだデータと、前記書き込みが行われた位置とを特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号とを第1の記憶サブシステムに送信するステップと、

第1の記憶サブシステムが前記データと前記シーケンス番号とを受信してこれらを第1の記憶資源に記憶し、前記シーケンス番号と、前記データが書き込まれた第1の記憶資源上の格納位置を特定する書き込み位置情報とを対応づけて記憶するステップと、

第2の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記シーケンス番号とを第3の記憶サブシステムに送信するステップと、

第3の記憶サブシステムが前記データと前記シーケンス番号とを受信して前記データを第3の記憶資源に記憶するとともに前記データと対になっていた前記シーケンス番号を第1の記憶サブシステムに送信するステップと、第1の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、第3の記憶資源に未反映となっているデータを把握するステップと、

を有することを特徴とするリモートコピー制御方法。

【請求項20】 前記第2の記憶サブシステムが障害等により使用できなくなった場合に、

前記第1の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第3の記憶資源において未反映となっている差分のデータ

とその書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第3の記憶資源に記憶して前記第1の記憶資源と前記第3の記憶資源の内容を同期させるステップと、

第1の記憶サブシステムが、前記ホストからの指示により前記第1の記憶資源に対してデータ書き込みが行われた場合に、前記データ書き込みにより書き込まれたデータと前記第1の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記データと前記書き込み位置情報とを受信して前記データを前記第3の記憶資源の前記書き込み位置情報で特定される位置に記憶するステップと、を有することを特徴とする請求項19に記載のリモートコピー制御方法。

【請求項21】 前記第2の記憶サブシステムが回復し再び使用できるようになった場合に、

前記第1の記憶サブシステムが、前記ホストにより書き込まれた前記第1の記憶資源に格納されている全てのデータを前記第2の記憶資源に転送するステップと、

第1の記憶サブシステムが、前記ホストからの指示により第1の記憶資源に対してデータ書き込みを行うステップと、

第1の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記シーケンス番号とを第2の記憶サブシステムおよび第3の記憶サブシステムに送信するステップと、

第2の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを第2の記憶資源に適時に記憶するステップと、

前記第2の記憶サブシステムが、前記第2の記憶資源に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶するステップと、

前記第3の記憶サブシステムが、前記データと前記シーケンス番号とを受信して前記データを前記第3の記憶資源に記憶するとともに前記対応づけにおける前記シーケンス番号を前記第2の記憶サブシステムに送信するステップと、

前記第1の記憶サブシステムから前記第2の記憶サブシステムへのデータの転送を停止するステップと、

前記第2の記憶サブシステムが、前記シーケンス番号を受信して、前記シーケンス番号と記憶している前記対応づけとに基づいて、前記第3の記憶資源に未反映となっているデータを把握するステップと、

前記第2の記憶サブシステムが、前記受信した前記シーケンス番号と前記対応づけとに基づいて把握した前記第

3の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第3の記憶サブシステムに送信するステップと、

前記第3の記憶サブシステムが、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第3の記憶資源に記憶して前記第1の記憶資源と前記第3の記憶資源の内容を同期させるステップと、

を有することを特徴とする請求項20に記載のリモートコピー制御方法。

【請求項22】 前記第1の記憶サブシステムと前記第2の記憶サブシステムとの間は同期転送により運用され、前記第1の記憶サブシステムと前記第3の記憶サブシステムとの間は非同期転送により運用されることを特徴とする請求項15～17および19～21のいずれかに記載のリモートコピー制御方法。

【請求項23】 前記第2の記憶サブシステムもしくは前記第3の記憶サブシステムは、前記第1の記憶サブシステムにおける書き込み位置情報に対応づけた前記第2の記憶資源もしくは前記第3の記憶資源における格納位置にデータを記憶することを特徴とする、請求項15～17および19～21のいずれかに記載のリモートコピー制御方法。

【請求項24】 ホストが接続された記憶サブシステムAとこれに通信手段を介して接続する複数の他の記憶サブシステムBと、を有する広域データストレージシステムにおけるリモートコピー制御方法であって、

記憶サブシステムBが、前記ホスト、記憶サブシステムA、前記通信手段のうち少なくともいずれか一つに障害が発生したことを検知するステップと、

記憶サブシステムBが、障害を検知した場合に記憶サブシステムBの中から記憶サブシステムAの処理を代行させる記憶サブシステムbを選出するステップと、

記憶サブシステムBが、選出した記憶サブシステムbとこれ以外の記憶サブシステムBが記憶しているデータの内容を一致させるステップと、記憶サブシステムBが、記憶サブシステムbに接続する副ホストにより前記ホストの運用を引き継ぐステップと、

を備えることを特徴とするリモートコピー制御方法。

【請求項25】 障害が発生したことを検知する前記ステップ、および記憶サブシステムAの処理を代行させる記憶サブシステムbを選出する前記ステップを、一台以上の前記記憶サブシステムBが行うことを特徴とする請求項24に記載のリモートコピー制御方法。

【請求項26】 前記障害が発生したことを検知するステップが、

前記記憶サブシステムAから前記記憶サブシステムBに対して送信されるハートビートメッセージがあらかじめ設定された時刻に受信できない場合に障害が発生したと認識するステップであることを特徴とする請求項24に

記載のリモートコピー制御方法。

【請求項27】 障害を検知した場合に記憶サブシステムBの中から記憶サブシステムAの処理を代行させる記憶サブシステムbを選出する前記ステップが、前記各記憶サブシステムBに記憶しているデータの更新順序を示すシーケンス番号のうち最新のシーケンス番号同士を比較して、これらのうちで最新の前記シーケンス番号を記憶している記憶サブシステムBを、前記記憶サブシステムAの処理を代行させる前記記憶サブシステムbとして選出するステップであることを特徴とする請求項24に記載のリモートコピー制御方法。

【請求項28】 前記各記憶サブシステムBに記憶しているデータの更新順序を示す前記シーケンス番号に抜けがある場合に、連続しているシーケンス番号のうち最新のシーケンス番号を用いて前記比較を行うことを特徴とする請求項27に記載のリモートコピー制御方法。

【請求項29】 選出した記憶サブシステムbとこれ以外の記憶サブシステムBが記憶しているデータの内容を一致させる前記ステップが、前記記憶サブシステムbとこれ以外の記憶サブシステムとの間で、全データの複写、もしくは、各記憶サブシステムBに記憶しているデータの差分データの複写により、前記選出した記憶サブシステムbとこれ以外の記憶サブシステムBが記憶しているデータの内容を一致させるステップであることを特徴とする請求項24に記載のリモートコピー制御方法。

【請求項30】 記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおける請求項14に記載のリモートコピー制御方法に使用する前記記憶サブシステムであって、前記記憶資源に対してデータ書き込みを行う手段と、前記記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶する手段と、

前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先に送信する手段と、前記転送先から送られてくる前記シーケンス番号を受信する手段と、記憶している前記対応づけと前記転送先から受信した前記シーケンス番号とに基づいて、前記転送先において未反映となっている書き込みデータを把握する手段と、を備えることを特徴とする記憶サブシステム。

【請求項31】 第1の記憶資源に対するデータ書き込み手段を備え第1のサイトに設置された第1の記憶サブシステムと、第1の記憶サブシステムに接続するホストと、第2の記憶資源に対するデータ書き込み手段を備え第2のサイトに設置された第2の記憶サブシステムと、第3の記憶資源に対するデータ書き込み手段を備え第3のサイトに設置された第3の記憶サブシステムと、を有

する広域データストレージシステムにおける請求項15に記載のリモートコピー制御方法に使用する前記第1の記憶サブシステムとして機能する記憶サブシステムであって、

前記ホストからの指示により第1の記憶資源に対してデータの書き込みを行う手段と、

第1の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶する手段と、

前記データ書き込みにより書き込んだデータと前記対応づけとを第2の記憶サブシステムに送信する手段と、第3の記憶サブシステムから送信されてくるシーケンス番号を受信して、このシーケンス番号と記憶している前記対応づけとに基づいて、第3の記憶資源に未反映となっているデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項32】 記憶資源に対するデータ書き込み手段を備え前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムAにおける請求項18に記載のリモートコピー制御方法に使用する前記転送先として機能する記憶サブシステムであって、記憶サブシステムAから送られてくる、前記記憶資源に対して書き込まれたデータと、そのデータの書き込みが行われた位置を特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号との対応づけとを受信して記憶する手段と、

別の前記転送先である転送先Bから送られてくる前記シーケンス番号を受信する手段と、

記憶している前記対応づけと転送先Bから受信した前記シーケンス番号とに基づいて、転送先Bにおいて未反映となっている書き込みデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項33】 第1の記憶資源に対するデータ書き込み手段を備え第1のサイトに設置された第1の記憶サブシステムと、第2の記憶資源に対するデータ書き込み手段を備え第2のサイトに設置された第2の記憶サブシステムと、第2の記憶サブシステムに接続するホストと、第3の記憶資源に対するデータ書き込み手段を備え第3のサイトに設置された第3の記憶サブシステムと、を有する広域データストレージシステムにおける請求項19に記載のリモートコピー制御方法に用いられる前記第1の記憶サブシステムとして機能する記憶サブシステムであって、

第2の記憶サブシステムから送信されてくる、前記ホストからの指示により書き込んだデータと、前記書き込みが行われた位置とを特定する書き込み位置情報とデータの書き込み順に付与されるシーケンス番号との対応づけとを受信して、これを第1の記憶資源に記憶する手段と、

第3の記憶サブシステムから送られてくるシーケンス番号を受信して、このシーケンス番号と、自身が記憶している前記対応づけとに基づいて、第3の記憶資源に未反映となっているデータを把握する手段と、

を備えることを特徴とする記憶サブシステム。

【請求項34】 ホストが接続された記憶サブシステムAとこれに通信手段を介して接続する複数の他の記憶サブシステムBと、を有する広域データストレージシステムにおける請求項24に記載のリモートコピー制御方法に使用する前記記憶サブシステムBとして機能する記憶サブシステムであって、

前記ホスト、記憶サブシステムA、前記通信手段のうち少なくともいずれか一

つに障害が発生したことを検知する手段と、

障害を検知した場合に記憶サブシステムBの中から記憶サブシステムAの処理を代行させる記憶サブシステムbを選出する手段と、

選出した記憶サブシステムbとこれ以外の記憶サブシステムBが記憶しているデータの内容を一致させる手段と、

記憶サブシステムbに接続する副ホストにより前記ホストの運用を引き継ぐ手段と、

を備えることを特徴とする記憶サブシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、災害による外部記憶装置の障害が生じた後に、速やかに、その障害から復旧可能な広域データストレージシステムに係り、特に、外部記憶装置が相互に100kmから数百km隔てて設置され、相補的な動作を行う3つ以上の外部記憶装置からなる広域データストレージシステムに関する。

【0002】

【従来の技術】本件の出願人による特開平11-338647号公報には、システムとデータの2重化を同期又は非同期にて行うことが開示されている。また、本件の出願人による特開2000-305856号公報には、非同期で遠隔地にデータのコピーを行う技術が開示されている。

【0003】このように、本件の出願人は、大型計算機システム、サーバー、ネットワーク上のパーソナルコンピュータ、その他の上位計算機システム（以下、ホストという。）から、データの順序を特定する特別な制御情報を受領することなく、そのデータを受け取った外部記憶装置（以下、記憶サブシステムという。）が、そのデータを、遠隔地に設置された第2の記憶サブシステムに対し、そのデータの順序性を常時保証しながら、非同期転送により第2の記憶サブシステムへ連続して間断なく書き込むという、非同期リモートコピーの技術を所有している。

【0004】また、同期転送の技術を用いてコピーを行

うときは、ホストとこれに接続された記憶サブシステムとの間のデータ更新処理と、この記憶サブシステムと付近地又は遠隔地に設置された記憶サブシステムとの間のコピー制御が連動するため、巨視的にみて常に2つの記憶サブシステム間でデータが一致しており、その書き込み順序性も同時に保証されている。尚、適当なデータ転送経路を選択すれば、2つの記憶サブシステムの距離が100kmを超える場合であっても、同期転送によるコピーが可能である。

【0005】昨今、データを安全に格納し保持することが重要であるという認識が高まっており、データストレージの市場では、ディザスタリカバリシステムを要請する声が多く聞かれる。従来のように、データ格納拠点を2つ設け、かかる2地点間を同期転送又は非同期転送で結ぶことは実現されている。しかし市場は、第3、第4のデータ格納拠点（以下、データセンタという。）を要求し、これらの間での完全な又は完全に近いディザスタリカバリシステムの構築を望んでいる。

【0006】その理由は、3拠点以上のデータセンタを設置しておけば、これらのうち1箇所が災害に見舞われても、引き続き発生する災害のリスクを軽減するために、残る複数のセンタ間でデータ冗長化の回復・維持が図れるであろうという期待にある。

【0007】従来の技術では、3以上のデータセンタを構築した場合に、ホストから受領する1/Oを唯一の記憶サブシステムの論理ボリュームで受領し、これを複数のデータセンタへリモートコピー技術を用いて転送する際の配慮が十分で無かった。例えば、一つのデータセンタが災害によりダウンした場合に、残る2以上のデータセンタ間で、データの順序性を保証した論理ボリュームを構築できるか、更新状態を引き継ぎデータの不整合を無くすることができるか、附近地又は遠隔地に対するコピーを可能とするシステムの再構築ができるかといった問題に関し、配慮が足りなかった。

【0008】

【発明が解決しようとする課題】災害はいつ発生するか不明なため、3以上のデータセンタ間で、常時、データ更新の順序性を保持しなければならない。

【0009】このため、ホストに特殊な機能を具備せず、複数のリモートコピー構成を連結し、同一論理ボリュームが受領したデータを、遠隔地又は附近地の別の記憶サブシステムへ配信し、かつ、如何なる時点で災害が発生しても、ホストからのデータの更新順序を、各データセンタの記憶サブシステムで、常時、保証する広域データストレージシステムを構成しなければならない。

【0010】本発明に係る広域データストレージシステムでは、記憶サブシステムの内部に、冗長化した論理ボリュームを設けることなく、別の記憶サブシステムに対し、データをコピーすることにより上記の課題を解決している。

【0011】また、本発明に係る広域データストレージシステムでは、災害後の復旧作業として、広域ストレージシステムの再構成を想定しており、正常な運用時に、直接、データ転送を行っていない記憶サブシステム間で、管理情報をやり取りし、データの更新状態を各記憶サブシステムで監視し管理する。そして、災害後の復旧作業（再同期、リシンク）において、災害発生直前に各記憶サブシステムが保持しているデータの差分のみを転送することで、即時に、ホストの交代（failover）と、アプリケーション実行の継続を行う。

【0012】＜データ更新の順序性を常時保証することについて＞ここで、順序性の保持の時間的範囲について補足説明する。

【0013】ホストから発行されたI/Oは記憶サブシステムに書き込まれ、記憶サブシステムが報告するデータの書き込み完了報告を認識し、ホストは次のステップを実行する。ホストは記憶サブシステムのデータ書き込み完了を受領しない場合又は障害報告があった場合は、次のI/Oを正常には発行しない。従ってデータの書き込みの順序性は、ホストが記憶サブシステムから書き込み完了報告を受領する前後で、記憶サブシステムが順序性保存の何らかの処理をすることで維持されるべきものである。

【0014】同期転送のリモートコピーでは転送されコピーされるデータが付近地又は遠隔地（以下、単に「別地」と略記する。）の記憶サブシステムに書き込まれ、別地の記憶サブシステムからの書き込み完了を受領した後、ホストに対し書き込み完了報告を行う。リモートコピーを行わない場合と比較し、リモートコピーに係る処理、及びデータ転送処理時間が長くなり、性能が遅延する。リモートコピーにおける接続距離を延長すると、データ転送に伴う処理時間が増大し、リモートコピーを行うことによりホストのI/O処理の性能をさらに低下させる。これを打破する一つの方法が非同期転送である。

【0015】非同期転送は、ホストからI/Oを受領した記憶サブシステムが、別地の記憶サブシステムへのデータ転送を行ない別地の記憶サブシステムの書き込み完了を待たずに、ホストからI/Oを受領した記憶サブシステムが書き込み完了報告をホストへ返す。これにより、記憶装置サブシステム間のデータ転送は、ホストのI/O処理と関係がなくなり、ホストのI/O処理と非同期に実行できる。しかし、ホストからのデータの到着順序を守って、別地の記憶サブシステムへデータを書き込まなければ、別地の記憶サブシステムのデータ順序性は維持されず、両記憶サブシステム間でデータの不整合を来す可能性がある。データの順序性を常時保証する機能を追加すれば、このような可能性を極小化できる。

【0016】別地の記憶サブシステムは、ホストI/Oを受領した記憶サブシステムと比較し、通常はデータの

更新は遅れているが、ホストからのデータ到着順序を守って記憶サブシステムへ書き込む限り、データの順序性に矛盾は無く、ジャーナルファイルシステムやデータベースリカバリ処理により、障害時の回復が可能である。

【0017】一方、データの順序性を維持せず、別地の記憶サブシステムへリモートコピーしてデータを反映させる方法もある。この方法は、ある時点までのホストから受領したデータを別地へ送り、それらを記憶サブシステムへ纏め書きする。ある時点までのデータ書き込みが終わった段階で、データ転送を終了し、以降、次の纏め書きまで、リモートコピーのデータ転送を抑止し、抑止している間のデータ順序性、ホストから受領したI/Oの一貫性を保証する。

【0018】この方法では、データの順序情報を付与する機能が不要であるが、ある程度の更新分のデータを蓄えておいて、その更新分を一括転送し、リモートへ書き込みが全て完了した段階で、データ整合性を保証している。この方法ではリモートコピーを行っている間に障害が発生すると、リモート側のデータ更新は順序性を維持して更新されていないため全滅となる恐れがある。リモートコピーのデータ転送を止めている間のみ、データ整合性を保証でき、adaptiveと呼ばれている。

【0019】出願人の所有する“データの順序性を常時保証する非同期転送によるリモートコピー”の技術によれば、ホストに完了報告を返す際に、記憶サブシステムがデータの順序性を保証する処理をしていることが特徴である。記憶サブシステムの制御装置におけるオーバーヘッドや内部処理の遅延時間に拘らず、ホストに返す際にデータ順序情報をブロック毎に管理する措置を施しているため、常時順序性を保証できる。

【0020】実際には、ホストから受領するI/O発行間隔よりかなり短い時間で、ブロック毎の制御・管理をしている。この一方で、リモート側の記憶サブシステムでデータ配信を待ちきれずタイムアウト（Timeout）とする値は、1時間以上に設定可能でもある。大切なのは、出願人のリモートコピーの技術が、データに順序情報を付与してデータブロックを転送し、これに基づきデータの順序を守って書き込みを行なっている点である。ローカル/リモートのデータ更新の時間差が、例えば半日あっても、順序性さえ正しければ、更新データ全てを喪失してしまう不整合より良いからである。

【0021】

【課題を解決するための手段】データを同期及び非同期に転送可能な転送経路、所定の管理情報のやり取りが可能な通信線路、及び、データ更新進捗管理手段により、3以上のデータセンタを相互に連結する構成とする。

【0022】ここで、更新進捗管理手段は、各記憶サブシステムに設けられ、いつ発生するか分からない災害に対応するため、他のデータセンタに設置された記憶サブシステムにおけるデータ更新の進捗状態を、適宜、監視

し、相互にその記憶サブシステムのデータ更新状態を把握させる手段である。

【0023】具体的には、直接データ転送を行なっていない記憶サブシステムの各々が、転送状態/ビットマップを持ち、転送ブロックのどの位置が何回更新されたか、一方が問合せ、他方がこれに回答することで、データ更新（リモートコピー）の進捗を監視し管理する機能を有する。

【0024】

【発明の実施の形態】3以上のデータセンタに、それぞれ設置された記憶サブシステムの間を、同期転送により連結する。

【0025】かつ、データの順序性を常時、連続的に保証する非同期リモートコピーの技術で連結する。そして、1箇所のプライマリデータセンタの記憶サブシステムから、これを除いた残りの別拠点の2以上のデータセンタの各記憶サブシステムへ、ホストからプライマリの記憶サブシステムが受領したデータを、ホストが更新した順序を保持しつつ、連続的に転送し格納する。

【0026】データが、ホストからの更新順を保証して冗長構成化されるため、万一、データセンタに災害・障害が発生しても、残ったデータセンタの記憶サブシステムの間で、各記憶サブシステム間の差分データのみを転送することで、即時に、リモートコピーの運用構成を回復でき、又は、データ喪失を最小限度とすることができる。

【0027】＜同期・非同期について＞まず始めに、図5、図6を用いて同期転送によるコピー又は非同期リモートコピーを定義する。

【0028】同期転送によるコピーとは、ホスト1から記憶サブシステム1に、データの更新(書き込み)指示があった場合に、その指示対象が附近地に設置された記憶サブシステム2にも書き込むデータであるときは、記憶サブシステム2に対して、指示された更新(書き込み)が終了してから、ホストに更新処理の完了を報告する処理手順をいう。ここで、附近地とは、いわゆるメトロポリタンネットワークと称される100km程度までの範囲を言うものとする。

【0029】つまり、同期転送のリモートコピー（図5）では、ホスト1から受領した更新データブロックを記憶サブシステム1で受領し（①）、そのデータブロックを記憶サブシステム2に転送し（②）、書き込み完了後、これを記憶サブシステム1で受領し（③）、最後にホスト1に対し更新データブロックの書き込み完了を行う（④）。途中の処理に失敗した場合には、ホスト1に書き込み障害を報告する。

【0030】同期転送によるコピーを実施すると、ホスト1に接続された近い方の記憶サブシステム1と、附近地に設置された遠い方の記憶サブシステム2のデータの内容が、巨視的にみて常に一致している。このため、災

害により一方がその機能を失った場合であっても、他方の記憶サブシステムに災害直前までの状態が完全に保存されているので、残るシステムで迅速に処理を再開できる効果がある。尚、巨視的にみて常に一致とは、同期転送の機能を実施中には、制御装置や電子回路の処理時間の単位（ μ sec、msec）で、一致していない状態が有り得るが、データ更新処理完了の時点ではデータは必ず同一の状態になっていることを意味している。これは、附近地の記憶サブシステムへの更新データの反映が終了しない限り、ホストに近い側の記憶サブシステム1の更新処理を完了できないためである。

【0031】一方、非同期リモートコピー（図6）とは、ホスト1からこれに接続された近い方の記憶サブシステム1に、データの更新(書き込み)指示があった場合、その指示対象が遠隔地に設置された記憶サブシステム2にも書き込むデータであっても、記憶サブシステム1の更新処理が終わり次第、ホスト1に対し更新処理の完了を報告し、遠隔地の記憶サブシステム2におけるデータの更新（反映）が、近い方の記憶サブシステム1における処理とは非同期に実行される処理手順をいう。

【0032】このため、近い方の記憶サブシステム1で必要とされる処理時間でデータ更新が終了するので、遠隔地の記憶サブシステム2へのデータの格納に起因する伝送時間、格納処理時間等により、ホスト1の処理が待たされることがない。ここで、遠隔地とは、いわゆるトランスコンチネンタルネットワークと称される、附近地より遠いが、距離の制約なく通信又はデータ転送可能な地点を言うものとする。

【0033】より具体的には、非同期リモートコピーでは、ホスト1から受領した更新データブロックを記憶サブシステム1で受領し（①）、ホスト1に対し更新データブロックの書き込み完了を行う（②）。記憶サブシステム1は、自己のスケジュールで、ホスト1の処理とは非同期に、記憶サブシステム2へデータを転送する。

【0034】遠隔地又は附近地へのデータ転送経路の複雑化、途中のデータ転送経路のボトルネックにより、データ転送中の当該データの順序性は保証されない（図6、点線の楕円内参照）。

【0035】一般に、データ転送の性能を上げるため、多くは高速転送のため、転送元から複数の転送経路を用いてデータを転送する場合がある。また、転送先まで遠距離となると、転送元は1つの転送経路であっても介在する交換機、ルーターその他の通信中継機器により、転送先まで転送経路が1本であることは保証されない。このように複数の転送経路を用いてデータを転送する場合には、経路によっては時間的な差異が生じ、遅い経路と早い経路とを介してデータが送られるため、転送先においてデータの順序が保存されないものである。

【0036】図6の楕円内に一例を示すが、データ転送経路上の順序を、Data #1、Data #2、Data

a # 4、Data # 3としている。記憶サブシステム2における更新順序はData # 1、Data # 2、Data # 3、Data # 4の順序である。記憶サブシステム2において、転送されてきたデータの順序をソートして正規の順序に並べ直しているからである。この更新処理の直後に不慮の災害が発生しても、データ更新の順序が守られているため、記憶サブシステム2のデータベースやジャーナルファイルシステムは回復処理を行うことができる。逆に、更新処理の直前に災害が発生したときは回復処理は不可能であるが、ホストへの応答とは非同期に、記憶サブシステム間で連続的に中断無くデータ転送処理を行うことで、データ不整合を極小化でき、巨視的に見て、常時、更新データの順序性を確保できる。

【0037】本実施の態様では、ホスト1からデータブロックを受領し、記憶サブシステム2へ転送する際に、ホストからのデータ更新順序を示すシーケンス番号情報をデータに付して転送している。このため、記憶サブシステム2で、シーケンス番号情報に基づくソート制御を行い、順序性を保証して、データの格納を完了できる。このような一連のデータ転送・処理に必要な処理時間の後は、データの順序性が遠隔地の記憶サブシステム2において保持されている。このように非同期のコピーを、これに固有なデータ処理を連続して行うこと（非同期リモートコピー）で、常時、データ更新の順序性を保証することができる。

【0038】非同期リモートコピーは、ホスト1の処理性能を落とさず、記憶サブシステム1及び2の間の距離を拡大できる特長があり、かつ、常時、データの順序性が保証されるため、広域データストレージシステムの利用者が自己の業務を遂行する上で、ほぼ任意の時点のデータベースやジャーナルファイルシステムの整合性を、遠隔地に設置された記憶サブシステムにおいて確保できる特長を有している。

【0039】<広域データストレージシステム、その1>図1に本発明の広域データストレージシステムの全体構成を示す。図9は本発明の別の広域データストレージシステムの全体構成を示す図である。図10は図1と図9の構成の組み合わせによる応用例を示す図である。

【0040】図1において3カ所のデータセンタに記憶サブシステムを設置する。各データセンタには複数の記憶サブシステムが設置されても良いし、それらがリモートコピー機能を伴った接続形態となっても良い。アプリケーションはデータセンタ1に接続されたホストで実行される。尚、ホストとデータセンタ1とのデータ転送経路は、ファイバーチャネル、メインフレームインタフェース、イーサネット（登録商標）LAN、公衆回線、インターネットその他専用回線である。

【0041】データセンタ1とデータセンタ2は近接地に存在し、同期転送によりデータ転送し得る構成である。データセンタ1とデータセンタ3は遠隔地に存在

し、これらの間は非同期リモートコピーの技術によりデータ転送し得る構成である。

【0042】正常な運用形態では、ホストからデータセンタ1が受領した更新データは、データセンタ1に設置された記憶サブシステムに格納され運用される。この更新データは、附近地に設置されたデータセンタ2の記憶サブシステムへ、ファイバーチャネル、メインフレームインタフェース、イーサネットLAN、公衆回線、インターネットその他専用回線を介して、同期転送される。つまり、データセンタ1とデータセンタ2では、記憶サブシステム間のデータ整合性は巨視的には絶えず保たれている。

【0043】正常な運用形態では、また、ホストからデータセンタ1が受領した上記の更新データを、遠隔地に設置されたデータセンタ3の記憶サブシステムへ、上記と同様な専用回線を介して、上記の同期転送処理と同時に、非同期リモートコピーの技術で転送される。尚、データセンタ1とデータセンタ2、データセンタ1とデータセンタ3、それぞれの間のデータ転送経路は同一種類の回線にする必要はない。

【0044】データセンタ1とデータセンタ3との間は遠距離であり、この間の転送に起因する更新データの到着順序の不整合が生じる。また、転送元となるデータセンタ1の記憶サブシステムには、転送先で未反映のデータとなる差分データが、存在することとなる。しかし、本発明の非同期リモートコピーでは、所定の非同期転送に固有のデータ処理後は、データベースやファイルシステムの回復処理に必要な、ホストからのデータ順序性を保証しているために、不整合を生じていたデータの順序を回復させることが可能である。この結果、データセンタ1とデータセンタ3の記憶サブシステム間では、上記ホストから受領した更新データの順序性は保たれる。

【0045】データセンタ2とデータセンタ3の間は、万一のリカバリ処理に備え、データを転送する通信線路は敷設・準備されているが、このストレージシステムの正常な運用時にはホストからの更新データは転送されない。データセンタ1での災害・障害発生の際に備え、正常な運用形態で、データ転送の進捗状態を問合せコマンドが、この通信線路を介して、データセンタ2からデータセンタ3へ、又は逆にデータセンタ3からデータセンタ2へ、送受信されることとなる。尚、敷設・準備された通信線路は、ファイバーチャネル、メインフレームインタフェース、イーサネットLAN、公衆回線、インターネットその他専用回線である。

【0046】正常時には、記憶サブシステム1と記憶サブシステム3との間の非同期リモートコピーにより為されたホストからの更新データの到着を、記憶サブシステム2から発せられる“データ進捗問合せコマンド”により、データセンタ2とデータセンタ3の間の通信線路を介して、問合せる。

【0047】“データ進捗問合せコマンド”の起動は、記憶サブシステム2のスケジュールに従って為される。記憶サブシステム1からの同期転送によるデータの受領のタイミングで、当該コマンドを発行しても良いし、所定の時間間隔で極めて問合せても良い。所定の時間間隔としては、例えば100msecから500msec毎に問合せても良いが、後述する転送状態/ビットマップの管理、これに基づく差分データの管理に時間が費やされ過ぎない程度となる。尚、1回の問合せで複数のビットマップを検査するようにしても良い。

【0048】正常な運用時には、記憶サブシステム2と記憶サブシステム3との間で、直接、データの転送は行なわれない。このため、記憶サブシステム2が“データ進捗問合せコマンド”を発行して、記憶サブシステム1と記憶サブシステム3のデータ更新状況を把握する。

【0049】万一、データセンタ1で障害が発生したときには、データセンタ2のホストを用いて、これまでのシステム運用を続行し（ホストのフェールオーバー）、記憶サブシステム2と記憶サブシステム3との間の差分データを、リカバリ処理に備えて敷設されたデータ転送の通信線路を用いてデータセンタ2からデータセンタ3へ転送する。差分データのための転送で即時に広域データストレージシステムを回復させることが可能である。尚、フェールオーバーとは、プライマリーシステムからサブシステムへ切り替えることをいい、古くは、ホットスタンバイとも呼ばれていた。

【0050】この後、データセンタ2からデータセンタ3へ、上記の通信線路を用いて、上述のような非同期リモートコピーを行うこととすれば、データセンタ1の復旧に伴う、データセンタ2とデータセンタ1との間の同期転送の復旧により、障害発生前の広域データストレージシステムを復旧させることができる。但し、障害発生前後で、データセンタ1とデータセンタ2の役割が入れ替わっている。

【0051】このように、附近地に存在する2つのデータセンタと、遠隔地に存在する2つのデータセンタとを統合し合計3つのデータセンタとすることで、リモートコピーの技術で連結する広域データストレージシステムとする。こうしておけば、中小規模の災害・障害のときは、附近地に存在する、相互に同期転送により連結されたデータセンタの一方で他方の代替を行うことができる。2つのデータセンタの記憶サブシステムのデータは同期転送により巨視的にみて一致しており、フェールオーバーが即時に行なえるからである。

【0052】<広域データストレージシステム、その2>図1のデータセンタ2とデータセンタ3との間の通信線路が非常用であるため、この通信線路を選択せず、障害・災害復旧後のデータセンタ1とデータセンタ3との間のデータ転送経路を選択する場合には、復旧後は、広域データストレージシステムは、図9の構成となる。

【0053】図9は、記憶サブシステム1と記憶サブシステム2が同期転送で、記憶サブシステム2と記憶サブシステム3が非同期リモートコピーで、それぞれ、接続された例である。図1の広域データストレージシステムにおいて、データセンタ1からデータセンタ2へ運用を切り替え、データセンタ2を主たる運用サイトとし、災害・障害復旧後は、データセンタ2からデータセンタ1へデータを同期転送させる一方で、データセンタ1からデータセンタ3へデータを非同期転送させる構成となるからである。

【0054】図9の場合において、直接データ転送に関与しない記憶サブシステム1から記憶サブシステム3へ、“データ転送進捗問い合わせ”コマンドが発行され、データセンタ3が応答して結果をデータセンタ1へ返す構成となっている。また図10は、図1と図9を組み合わせた構成である。記憶サブシステム3と5との間、記憶サブシステム2と5との間が、“データ進捗問い合わせコマンド”の発行・応答の経路に該当する。

【0055】上記の広域データストレージシステムの態様であれば、大規模な災害や、附近地に存在する2つのデータセンタに相次いで障害が発生した場合であっても、データセンタ3のホストへフェールオーバーすることで、災害直前のシステムが運用してきたデータを引き継いで処理でき、また、データの喪失を最小限度とすることができる。

【0056】つまり、附近地にある2つのデータセンタが全滅する程度の災害が発生したときは、遠隔地に存在するデータセンタ3又は5（図1、図9、図10）の記憶サブシステムを生かすことができる。ホストからの更新データの順序性が確保されつつ、非同期リモートコピーが行なわれているからである。但し、災害による未反映のデータは復旧できない。

【0057】<記憶サブシステムの構成について>図1、図9及び図10では、同期転送によるコピー及び非同期リモートコピーの組み合わせを示している。本来、リモートコピーは、1論理ボリュームと1論理ボリュームをデータ転送技術で結合したものである。本発明では、1個の論理ボリュームに対するデータ受領を、同期転送し、更に非同期転送して、附近地と遠隔地の双方にリモートコピー機能でデータ送信制御を行なっている。

【0058】これらは記憶サブシステムの制御装置のマイクロコードで実現される機能である。ホストや別の記憶サブシステムからの更新データは、一旦、キャッシュ5（図2）に格納される。この時点では、当該データは、まだ記憶サブシステム内のハードディスクドライブにRAID制御により書き込まれていない。キャッシュ5内で当該データの転送制御情報に加え、別の記憶サブシステムへリモートコピー転送したり、複数の記憶サブシステムとのリモートコピー構成を同時に実現する制御を行う。同期転送と非同期転送による組合せを守ること

により、いつ災害が発生しても、各データセンタでは、データの更新順序を保った、データベースやジャーナルファイルシステムがリカバリ可能な論理ボリュームを常時、保持していることとなる。

【0059】図2は、記憶サブシステムの概略構成を示す図である。

【0060】制御装置1は、ホスト及びリモートコピーの接続先とデータの送受を行うチャネルアダプタ3、ディスク装置2内のハードディスクドライブ7をディスクインタフェース8（ディスクI/F8）を介して制御するディスクアダプタ9を有する。

【0061】チャネルアダプタ3とディスクアダプタ9は、それぞれ、マイクロプロセッサを有し、データ転送バス11・制御バス11を介してキャッシュメモリ5と接続されている。尚、バス構成は一例であり、必要に応じてクロスバ構成としても良い。また、制御装置1を複数設けてクラスタ構成とし、複数の制御装置1を連絡する共通の第3のバスを追加しても良い。

【0062】ホストとの間や、リモートコピーの接続先とデータ送受を行う際の格納元は、キャッシュ5である。制御情報、構成管理情報、転送状態/ビットマップは、制御メモリ6に格納されている。

【0063】リモートコピーには送信及び受信の機能があり、本実施例ではホストからI/Oを受領するチャネルアダプタを分けて搭載している。ホストから受領したI/Oは、一旦、キャッシュ5へ格納される。リモートコピーの転送先情報や後述する状態管理/ビットマップは、制御データとして制御メモリ6に格納され、マイクロコードにより制御される。

【0064】キャッシュに受領したデータは、ディスクアダプタ9によりハードディスクドライブ7へRAID制御で書き込まれる。これとは別の処理である、マイクロコードを用いた制御により、予め定義されたリモートコピー転送先への送信制御が行なわれる。

【0065】例えば、ホストから受領したデータが、後続するリモートコピーの対象であり、非同期転送によるデータ送信を行うと定義されていた場合には、キャッシュ5の内部のデータに対して、データ受領順にシーケンス番号を付与する。これはデータ更新を示すID情報でもある。シーケンス番号を付与されたデータは、チャネルアダプタ3のリモートコピー送信機能により、当該シーケンス番号と共に送信される。

【0066】別の実施例で、ホストから受領した更新ブロックを、複数の論理ボリュームと接続するリモートコピー制御が定義されていた場合には、キャッシュメモリ5の内部のデータは、同期転送用に加工されると同時に、非同期転送用にも加工され、シーケンス番号が付与されて、それぞれ、チャネルアダプタ3で附近地又は、遠隔地に向けて送信される。

【0067】図2は本発明を実現する一例であり、本発

明はハードウェア構成に依存しない。リモートコピー接続が記憶サブシステム間で実現可能であれば、マイクロプロセッサによる論理的なサポート、マイクロコード制御で実現できるためである。

【0068】＜転送状態/ビットマップ＞図4は、転送状態/ビットマップ（適宜、ビットマップと略記する。）の一例を示したものである。これは、直接データ転送を行っていない2つのデータセンタに設置された記憶サブシステムの内部に、災害・障害の復旧の際にペアを組んであろう相手（別のデータセンタに設置された記憶サブシステム）のデータ更新の進捗状況を知るために用意されたものである。例えば、図1ではデータセンタ2とデータセンタ3との間で、非常時のためにペアが組まれる。図9の広域データストレージシステムであれば、記憶サブシステム1と記憶サブシステム3との間で、図10では、記憶サブシステム2と記憶サブシステム5、記憶サブシステム3と記憶サブシステム5との間で、それぞれ、非常時のために、ペアが組まれることとなる。

【0069】転送状態/ビットマップは、ペア（対）となる論理ボリュームに対して必要であり、本発明では1個の論理ボリュームの実体に対し、2個以上の転送状態/ビットマップを持ち得る。各ビットマップは、ペアやペアとなる場合を想定した定義付けにより、相手の論理ボリュームとの差分管理を行うために使われる。ビットマップの中のブロックナンバは、論理ボリュームの更新を管理する最小単位であるブロックに対応させた番号である。

【0070】ホストI/Oは、このブロックナンバと同一単位である必要はない。ホストI/Oの単位は、通常、最小で512バイトとされ上限も設けられているが可変である。一方、ビットマップは、50kB程度の大きさ、又は700kB程度の大きさのものもあるが、20kBから1000kB程度まで種々の大きさがある。ホストI/Oの1ブロックに対して、必ずしも、1ビットマップが対応する訳ではない。

【0071】ブロックナンバに対応するブロックの内容が更新されれば、差分管理は当該ブロックナンバ全体となり、同期（リシンク）を行うときに当該ブロックナンバのデータ全体が転送されることとなる。

【0072】ビットマップは、ブロックナンバ毎に、当該論理ボリュームの更新された単位として、リモートコピーによるペアを再構築する際（再同期、リシンク）に当該更新されたブロックのみを転送する目的で、相手論理ボリュームに転送すべき“Update”情報を持つ。つまりUpdateフラグがOn（図4の実施例では1）、であれば転送対象であることを示す。通常のUpdateは、ホストからのコマンド単位で為されることから、カウンタ値が0であることに基づき、Updateフラグを0とする。

【0073】ビットマップは更に、同一ブロックナンバーで、複数回の更新を記録するカウンタ値を持つ。カウンタ値は、更新が無ければ“0”、3回更新されれば“3”となる。ブロックナンバーで表されるデータブロックの大きさが、ホストから更新されるデータブロックより大きい場合には、このカウンタ値を使うことにより確実に相手論理ボリュームへ更新データのみを転送できることとなる。

【0074】後述の“データ進捗問合せコマンド”の中に格納されたブロックナンバーとカウンタ値と、問合せ先の記憶サブシステムのビットマップのブロックナンバーとカウンタ値との比較を、データコピー監視機能（後述）で行う。この際に、ある記憶サブシステムが持つカウンタ値が、この記憶サブシステムに送付されて来た“データ進捗問合せコマンド”に記述されたカウンタ値と等しいか大きい場合に、所定の記憶サブシステムのビットマップのカウンタ値は1減算される処理を受ける。

【0075】送付されて来た“データ進捗問合せコマンド”に記述されたカウンタ値未満である場合は、その記憶サブシステムのビットマップのカウンタ値は何ら処理を受けない。そして減算したかしないかを、“データ進捗問合せコマンド”に回答して返す。

【0076】その記憶サブシステムのビットマップのカウンタ値が、送付されて来た“データ進捗問合せコマンド”に記述されたカウンタ値“以上”の場合には、データ更新の進捗は、正常なリモートコピー機能により既に、その記憶サブシステムにおいて格納済み、書き込み済みであることを意味する。また“未満”の場合には、データが未到着であることを意味している。

【0077】図4のカウンタ値は有限であり、例えば、1バイト分をカウンタ値として割り当てた場合には、256回を超える管理はできない。この例では同一ブロックが256回を超えた更新を受けた場合には、最早、カウンタ値のUpを行わず、Updateフラグを恒久的に立ててしまう処理を行う。つまり図4でカウンタ値に“Over Flow”を意味する情報を格納する。

【0078】このような恒久的な指定がなされると（図4、Over Flow）、ビットマップで特定される、恒久的指定のなされたブロックのUpdateフラグの解除（0を入力すること）は、相手論理ボリュームへの転送が完了しコピーが確定したことを、このビットマップを有する記憶サブシステムが認識するときまで行わない。

【0079】カウンタ値を用いた更新管理を行う理由を次に補足説明する。

【0080】例えば、50kB程度のデータ量を有するトラックに対応させてビットマップの管理を行う場合には、この50kBのデータのうち、異なる3箇所が、異なる時刻において、それぞれ更新されたとする。トラックに対応させてビットマップ管理を行うのは、災害・障

害後の復旧（再同期、リシンク）において扱う単位がトラック単位であるためである。

【0081】カウンタ値による管理を行わない場合には、Updateフラグのみ監視することとなるが、ある時刻でUpdateフラグが1であることのみを確認しても、その後の時刻に2度目、3度目の更新があった場合には、2度目以降のデータ更新を見逃してしまう。新たにカウンタ値の概念を導入して、ホストからのコマンド単位で為される同一データブロック（ここではトラックの一部）の更新を微細に監視することで、かかる不都合を防ぐことができる。

【0082】次に、図2の制御装置1の内部でマイクロコードにより実現される転送状態／ビットマップの機能について定義する。論理ボリュームはリモートコピーの対となる論理ボリュームとの間で下記の転送状態を有する。これらは同期転送又は非同期転送に依存しない。

【0083】1）「正常ベア状態」とは、データの順序性を保証して、双方のボリューム間で、同一のデータを2重に保持している状態をいう。

【0084】2）「転送抑止ビットマップ登録の状態」とは、データの更新をビットマップに登録する状態をいう。未だベアの相手へデータの転送は行なわれていない。

【0085】3）「ビットマップ使用のコピー状態」とは、「転送抑止ビットマップ登録の状態」から「正常ベア状態」への移行期をいう。2重化のためのコピーの初期状態に当たる。

【0086】4）「障害状態」とは、障害によりデータを転送できない状態をいう。ビットマップに登録される。

【0087】5）「ベア無ビットマップ登録状態」とは、本発明固有の特殊な状態をいう。災害・障害前に、相互にデータ更新状態を監視し保持する必要から生じた状態である。

【0088】6）「ベア無状態」とは、ビットマップは用意されているが、未だベアを組んでおらず、データ更新の情報が登録されていない状態をいう。

【0089】「ベア無ビットマップ登録状態」が存在することが本発明の特徴となる。この状態を持つことなく、“転送抑止ビットマップ登録の状態”というサスペンド（Suspend）状態で兼ねても良い。ここで、サスペンド状態とは、論理ボリュームへのデータの更新状態を、ビットマップでのみ管理し、リモートコピーによる転送制御を行なわない状態をいう。

【0090】「ベア無ビットマップ登録状態」を持つのは、転送状態／ビットマップをベアで持つ必要からである（図3）。例えば、図1の広域データストレージシステムにおいては次の理由による。

【0091】データセンタ3が保持するデータを監視するため、データセンタ2の記憶サブシステムの内部の論

理ボリュームに対応して設けられた転送状態/ビットマップに、データセンタ3のデータ更新状態を持つ必要があり、且つ、データセンタ2が保持するデータを監視するため、データセンタ3の記憶サブシステムの内部の論理ボリュームに対応して設けられた転送状態/ビットマップに、データセンタ2のデータ更新状態を持つ必要があるためである。

【0092】図9の広域データストレージシステムにおいては、データセンタ2の障害発生に備えて、データセンタ1とデータセンタ3のリモートコピーの差分管理情報から、データセンタ1とデータセンタ3との間でペア構築を目的として、「ペア無ビットマップ登録状態」をデータセンタ1とデータセンタ3で持つ必要がある。この結果、記憶サブシステムやデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、障害回復後に更新部分のみの差分転送が可能となる。

【0093】転送状態/ビットマップの機能は、上記の様な制御を実現するマイクロコード及びビットマップと関連する制御テーブルから成る。具体的機能は、例えば、図2のマイクロプロセッサ4のマイクロコードと制御メモリ6で行なわれるが、先に示した様にマイクロコードの制御により自由に実装できる。例えば、マイクロプロセッサ10による実現も可能である。またマイクロプロセッサが1台のみの制御装置でも実現できる。

【0094】<広域データストレージシステムの運用>図3は、図1の広域データストレージシステムが正常に運用されている場合の基本的な制御方法を説明するための概略図である。正常運転ではデータ進捗問合せコマンドを記憶サブシステム2から記憶サブシステム3へ送信する。例えば、記憶サブシステム1の障害の際、実際の差分データの転送に際しては、記憶サブシステム2と記憶サブシステム3との間で、転送状態/ビットマップの機能を使用し、両方の記憶サブシステムのビットマップについて、論理演算を行う。その結果に基づき、相当するデータブロックのみを記憶サブシステム2から記憶サブシステム3へ転送している。図8に、図1の広域データストレージシステムのデータセンタ1に障害・災害が発生した場合において、非同期リモートコピーを再開させる概略の手順を示す。

【0095】図8において、正常な運用では、データセンタ1から附近地のデータセンタ2へ同期転送によりデータの二重化が図られる一方で、遠隔地のデータセンタ3へは非同期転送によりデータの更新順序を確保したコピーが行なわれている。そして、データセンタ2の記憶サブシステム2のスケジュールで、データ進捗問合せコマンドがデータセンタ3に対し発行され、データセンタ2と3とは管理情報をやり取りして、データの差分管理を行なっている。

【0096】データセンタ1に災害・障害が発生する

と、データセンタ2の記憶サブシステムは、非同期転送により、差分データをデータセンタ3へ送付し、即時に、データセンタ2と遠隔地のデータセンタ3によるシステム運用を回復できる。

【0097】図3において、転送状態/ビットマップは、1論理ボリューム当たり2個持ち、それぞれが、これらのビットマップを用いた機能を有する。記憶サブシステム1は、記憶サブシステム2と記憶サブシステム3に対し、転送状態/ビットマップ#1に対応する機能及びビットマップ#2に対応する機能を持つ。

【0098】記憶サブシステム2と記憶サブシステム3は、同期転送及び非同期転送の各々について転送状態/ビットマップ#3及び#6の機能をそれぞれ持つ。これら#1と#3、#2と#6のそれぞれの機能は、正常運転の際には、「正常ペア状態」を格納している。

【0099】転送状態/ビットマップ#4及び#5の機能は、それぞれ、記憶サブシステム2及び記憶サブシステム3が持っている。この広域データストレージシステムが正常に運用されているときには、転送状態/ビットマップ#4及び#5の機能は、上述の「ペア無ビットマップ登録」状態を保持する。

【0100】転送状態/ビットマップ#4の機能は、記憶サブシステム3の論理ボリュームに対する差分管理を、転送状態/ビットマップ#5の機能は、記憶サブシステム2の論理ボリュームに対する差分管理を、それぞれ行う。

【0101】図10の拡張として、ホストからのI/Oを受領する、第1のデータセンタに設置された記憶サブシステムの制御装置1が、N台の同期転送のコピー先と、M台の非同期リモートコピーのコピー先を持つ構成では、その制御装置1は、N+M個の転送状態/ビットマップの機能を有する。これに対応する、遠隔地又は附近地の記憶サブシステム(コピー先)も、転送状態/ビットマップを持つこととなる。この結果、制御装置1やデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、災害回復の際の更新部分のみの差分転送が可能となる。

【0102】<データコピー監視機能>次に、データコピー監視機能について説明する。この機能には、ビットマップの制御機能、リモートコピーのステータス管理機能、構成管理機能、データ進捗問合せコマンドの制御機能、リモートコピーのデータ転送指示機能等が含まれる。

【0103】図3の記憶サブシステム2の制御装置で、同期転送によるデータブロックを記憶サブシステム1から受領する。本データは記憶サブシステム2のキャッシュメモリに格納されディスクドライブで記憶される。この際、転送状態/ビットマップ#4の機能により当該データブロックが登録される。図4のビットマップに登録

する。

【0104】次に当該ブロックナンバとカウンタ値を格納した“データ進捗問い合わせ”コマンドを記憶サブシステム2から記憶サブシステム3に対して発行する。発行のタイミングは同期転送に基づいても良いし、記憶サブシステム2の独自のスケジュールで行なっても良い。

【0105】記憶サブシステム3の制御装置で、記憶サブシステム2からの“データ進捗問い合わせ”コマンドを受領し、転送状態/ビットマップ#4のブロックナンバとカウンタ値を切り出し、記憶サブシステム3の該当する転送状態/ビットマップ#5のそれらと比較する。

【0106】その結果、転送状態/ビットマップ#5のブロックナンバがUpdateフラグ1（更新）を示し、かつ、カウンタ値が転送されて来たもの以上であれば、同期転送に係るデータと、非同期リモートコピーに係るデータとが一致しているので、転送状態/ビットマップ#6の対応するブロックナンバから、カウンタ値を1減算する。

【0107】減算の結果、カウンタ値が“0”となった場合には、Updateフラグを“0”とする。カウンタ値が“Over Flow”である場合には、何も操作しない。

【0108】また、転送状態/ビットマップ#5に登録されていたカウンタ値が、記憶サブシステム2からの問合せコマンドから抽出されたカウンタ値未満であったり、Updateフラグが“0”（Off）で更新が示されなかった場合には、#5への更新は行わず、これをデータ進捗問合せコマンドの結果として記憶サブシステム2へ返す。

【0109】#5の転送状態/ビットマップの機能が、#6の転送状態/ビットマップのカウンタ値を減算するということは、記憶サブシステム1から同期転送により既に記憶サブシステム2へ到着したデータブロックが、記憶サブシステム1から記憶サブシステム3へ非同期転送により到着済であったことを意味している。

【0110】データコピー監視機能は、本応答結果を用いて記憶サブシステム2の転送状態/ビットマップ機能の制御を行なう。記憶サブシステム3で“データ進捗問い合わせ”コマンドのブロックナンバとカウンタ値が既に登録されていた旨の応答を返す場合（減算できた場合）には、記憶サブシステム2の制御装置でも転送状態/ビットマップの機能でカウンタ値の減算、Updateフラグの操作を同様に行う。

【0111】当該コマンドの応答結果が、未登録であれば、記憶サブシステム1から記憶サブシステム3へのデータの非同期転送が未完であるとして、記憶サブシステム2の転送状態/ビットマップ#4の機能は、自己のビットマップに更新状況を保持する。これは後に更新差分部分のみを再同期させる際の対象となる。

【0112】この時点で記憶サブシステム1が重大障害

を持ち、記憶サブシステム2と記憶サブシステム3との間でリモートコピー構成を再構築（再同期、リシンク）しなければならない場合には、ビットマップを参照した結果、未転送のデータのみ、即ち、差分のデータブロックのみを、記憶サブシステム2から3へ転送すれば良い。その結果、差分データの転送だけで即時に“正常ベア”を構築できる。これを実現する機能を“データコピー監視機能”と呼ぶ。

【0113】＜正常な運用の際に、直接データ転送を行なわない記憶サブシステム間での差分管理方法、その1＞図9の広域データストレージシステムにおいて、記憶サブシステム2に障害が発生した場合に、記憶サブシステム1と記憶サブシステム3との間で非同期リモートコピーによるシステム運用の復旧を図るときを考える。

【0114】このために、ホストからデータ更新を受領した記憶サブシステム1の制御装置1（図2）は、記憶サブシステム2の制御装置1の論理ボリュームに同期転送のコピーによるデータ転送を行う際に次の処理を行なう。

【0115】転送するブロックの位置情報を、記憶サブシステム1の制御装置1に存在するビットマップに、記憶サブシステム3の論理ボリュームの更新情報を格納する。このとき既に転送したブロックが記憶サブシステム3において更新されていたときは、ビットマップのカウンタ値を1増加（インクリメント）する。

【0116】記憶サブシステム1の制御装置1は、記憶サブシステム2の制御装置1に対して同期転送が完了した後、記憶サブシステム3の制御装置1に対して、同期転送したデータブロックが、記憶サブシステム2の制御装置1を経由して到着したか否かを問合せするため、記憶サブシステム1と記憶サブシステム3とを結ぶ通信線路を用いて、確認コマンドを発行する。

【0117】確認コマンドには、ホストから受領した更新データの記憶サブシステムにおけるデータブロックのブロックナンバとカウンタ値が含まれている。確認コマンドを受領した記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1経由で既に存在するデータブロックが、確認コマンドで問合せされたブロックと一致するか否かを判定する。

【0118】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1の論理ボリュームに対する転送状態/ビットマップの機能の他に、記憶サブシステム1の制御装置1の論理ボリュームに対する状態管理/ビットマップの機能を持つ。

【0119】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1からデータを受領すると、記憶サブシステム1の制御装置1の状態を格納すべく、自己の持つ転送状態/ビットマップへ登録する。このビットマップでは、論理ボリューム内のアドレスに関わるブロック位置に対する更新情報を有し、さらに複数回の同

一ブロックに対する更新を管理するためにカウンタ値を有している。

【0120】記憶サブシステム3の制御装置1の転送状態/ビットマップに登録された結果は、記憶サブシステム1の制御装置1から発行された確認コマンドのブロックナンバ及びカウンタ値と比較される。比較の結果、一致又はカウンタ値が確認コマンドにあったカウンタ値以上の場合には、データの到着が正常に完了していると判断し、転送状態/ビットマップの機能を用いてビットマップのカウンタを1減算する。

【0121】他方、記憶サブシステム1の制御装置1は、記憶サブシステム3の制御装置1から返される結果が、記憶サブシステム3へデータブロックが記憶サブシステム2を経由して到着していることを示す場合には、上述の記憶サブシステム3の制御装置が為したように、転送状態/ビットマップの機能を用いてビットマップのカウンタを1減算する。

【0122】以上のように、ビットマップを監視・管理することで、記憶サブシステム2が災害等により重大障害を持ち、同期及び非同期転送によるデータ送受が行えなくなった場合であっても、ホストがI/Oを発行する記憶サブシステム1と、記憶サブシステム2の内容を非同期リモートコピーにより格納した記憶サブシステム3との間で、非同期リモートコピーを構成することができる。

【0123】この際に記憶サブシステム1と3の、それぞれの制御装置の転送状態/ビットマップの機能により、論理ボリュームの全データをコピーすることなく、差分データのブロックのみを転送することにより、即時に、構築することができる。

【0124】＜正常な運用の際に、直接データ転送を行わない記憶サブシステム間での差分管理方法、その2＞図1の広域データストレージシステムにおいて、ペアになっている論理ボリューム間、つまり、記憶サブシステム1と2並びに記憶サブシステム1と3、それぞれの間のデータ更新状態の管理のために、転送状態/ビットマップの機能が各論理ボリューム毎に用意される。

【0125】記憶サブシステム1の制御装置1で障害が発生し、同期転送のコピー及び非同期リモートコピーの双方が継続不可となった場合には、記憶サブシステム2と3のそれぞれの制御装置1の間で、先ず差分データをコピーし両者を一致させる。次いで、記憶サブシステム2と3の間で非同期リモートコピーを構成する。

【0126】ホストから更新すべきデータを受領した記憶サブシステム1の制御装置1は、記憶サブシステム2の制御装置1へ同期転送によりデータブロックを送出し、これを記憶サブシステム2の制御装置1が受領する。記憶サブシステム2の制御装置1は、受領したデータブロックの位置情報（ブロックナンバ）を、記憶サブシステム3の制御装置1の配下の論理ボリュームの管理

情報との比較のために、自己が保持する転送状態/ビットマップに格納する。転送状態/ビットマップは、受領したデータブロックが更新された場合には、カウンタ値を1増加（インクリメント）する機能を備え、複数回のデータブロックの更新を記録できる。

【0127】記憶サブシステム2の制御装置1は、上記の転送状態/ビットマップへ所定の管理情報を登録した後、記憶サブシステム2の制御装置1と記憶サブシステム3の制御装置1との間を結ぶデータ転送経路を用いて、データブロックが記憶サブシステム3へ到着したか否かを問合せる確認コマンドを、記憶サブシステム3の制御装置1へ発行する。

【0128】確認コマンドは、記憶サブシステム2の制御装置1が、同期転送により記憶サブシステム1から受領したデータブロックの位置情報であるブロックナンバと、データブロックが何回更新されたかを示すカウンタ値を含む。

【0129】記憶サブシステム3の制御装置1は、記憶サブシステム1の制御装置1から、非同期リモートコピーの技術で受領したデータブロックの位置情報（ブロックナンバ）とカウンタ値を、記憶サブシステム2の制御装置1の配下の論理ボリュームの管理情報との比較のために、自己の制御装置1が持つ転送状態/ビットマップの機能を用いてビットマップに格納する。記憶サブシステム3の制御装置1は、ビットマップと確認コマンドの有する対応する値との比較を行う。

【0130】記憶サブシステム2から3へ問合せた確認コマンドが有するブロックナンバとカウンタ値と、記憶サブシステム3の制御装置1が持つ、記憶サブシステム2の制御装置1の配下の論理ボリュームの管理情報である、これらの値とを比較して、確認コマンドの値と同一又はカウンタ値が確認コマンドのカウンタ値より大きい場合には、転送状態/ビットマップの機能で、そのデータブロックのカウンタ値を1減算する。

【0131】減算した結果が0になる場合は、記憶サブシステム2と3との差分データはないことになるので、ビットマップの管理から削除する。上記の比較の結果が一致しない場合には、記憶サブシステム3の制御装置1は、ビットマップのカウンタ値を操作しない。

【0132】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1に、確認コマンドの応答である判定結果を返す。この結果を記憶サブシステム2の制御装置1が参照し、比較がカウンタ値を減算した場合には、既に記憶サブシステム2と3の間で、同一のデータブロックの更新が正常に終了していると断定する。

【0133】記憶サブシステム3に更新すべきデータブロックが届いていない場合には、記憶サブシステム2にのみ、更新に係るデータブロックが格納されていることになる。記憶サブシステム2の制御装置1は、自己の転送状態/ビットマップの機能でこれを記憶する。

【0134】記憶サブシステム2の制御装置1が、記憶サブシステム3の制御装置1から確認コマンドの応答を受領し、記憶サブシステム3に更新すべきデータブロックが未到着であった場合には、記憶サブシステム2の制御装置1が持つ、記憶サブシステム3の論理ボリュームの更新状態に対応する転送状態/ビットマップのカウンタ値は減算しない。このことは、そのビットマップは、更新に係るデータブロックが、記憶サブシステム2と3との間で差分であることを示す。

【0135】他方、データの到着完了を示した場合には、上記の転送状態/ビットマップの更新に係るデータブロックのカウンタ値を1減算する。カウンタ値が0のときは、記憶サブシステム2と3との間で、更新に係るデータブロックは同一であり不整合がないので、差分データのコピーの対象とはしない。

【0136】このように、正常運用の際に、直接データ転送を行っていない記憶サブシステムの制御装置同士が、災害・障害からの回復を想定して、論理ボリューム間の差分データ管理を行っているため、記憶サブシステム間で差分データのみをコピーし不一致をなくすることが高速に行なえる。

【0137】＜フェールオーバー後のシステムの運用＞図7に、図1の広域データストレージシステムが、フェールオーバーにより状態を遷移して図9の構成となった場合の運用について簡単に説明する。図3で記憶サブシステム1に、図9で記憶サブシステム2に、図10で記憶サブシステム1、2又は4に、それぞれ重大な障害が起きた場合には、図7に示す様に、残存する2以上の記憶サブシステムの間で、リモートコピー構成の復帰を図ることとなる。

【0138】本発明によれば、図7の様に、直接データ転送に関与していなかった論理ボリューム間（記憶サブシステム1と記憶サブシステム3との間）で、差分データのみコピーすれば、即時に、リモートコピーのペアを生成でき、リモートコピーの運用再開が可能である。

【0139】本発明を実施しない場合には、図3の記憶サブシステム2と3の間、図9の記憶サブシステム1と3の間で、それぞれ、リモートコピー構成をつくるに際し、図3の構成では記憶サブシステム2から記憶サブシステム3に対し、図9の構成では、記憶サブシステム1から記憶サブシステム3に対し、それぞれ、記憶サブシステムが保持するデータのフルコピーを行なわなければならない。大規模のデータセンタでは、コピーに長時間を要し、リモートコピーの運用再開が遅くなる。長時間を要するコピー中に、再度、コピー元やデータ転送経路に障害・災害が発生すると、データは破壊され喪失することとなる。

【0140】図11を用いて、図9の構成におけるデータコピー監視機能について簡単に説明する。

【0141】データ進捗問合せコマンドは記憶サブシス

テム1から記憶サブシステム3に対して発行される。データコピー監視機能は図1の場合と一部処理が異なる。記憶サブシステム1が、同期転送により記憶サブシステム2へホストから受領した更新データを転送した後、記憶サブシステム1から3に対し、上述した“データコピー監視機能”を作動させる。つまり、“データ進捗問い合わせ”コマンドを発行し、記憶サブシステム1の持つ転送状態/ビットマップ#1と、記憶サブシステム3の持つ転送状態/ビットマップ#3で、それぞれのUpdateフラグ、カウンタ値を登録し、所定の操作を行う。

【0142】記憶サブシステム1から3に、ホストから記憶サブシステム1が受領したデータ（トラック）と同じデータが、記憶サブシステム3に届いたか否か、問合せた結果、未着であれば、記憶サブシステム1の転送状態/ビットマップ#1のビットマップは、そのまま保持する。結果が到着であれば、つまり、#3のビットマップのブロックナンバ、カウンタ値が同一であれば、Updateフラグを削除し、#1のビットマップを削除する。

【0143】＜再同期におけるその他の処理＞データコピー監視機能で検出した“データ進捗問い合わせ”コマンドの応答結果に、エラーや不具合（タイムアウト）が生じたり、転送状態/ビットマップの機能に不具合が生じた場合には、障害・災害の際に行われるべき回復処理に関する差分管理を禁止する。

【0144】転送状態/ビットマップの機能において、ビットマップは有限なカウンタ値の格納領域を有している。この有限値を超えて（オーバーフロー）、同一データブロックが更新された場合には、そのデータブロックは、その後2以上の記憶サブシステム間で冗長度が維持されていても、災害・障害発生後に再同期処理、差分コピー処理が行なわれる際に、必ず更新対象として扱う。

【0145】正常な運用において直接、データ転送を行なわない記憶サブシステム間で取り取りされる問合せ（確認コマンド送出）に対し、所定時間、応答が無い場合は、タイムアウトであるとして再同期処理を禁止する。非同期リモートコピーによるペアの再構築処理や、差分データのみ転送する処理を行わず、禁止する。ペアの相手のデータ更新状態を知ることができないため、そのままペアの再度構築処理を行なわしめることは妥当でないからである。

【0146】＜非同期転送におけるデータの整合性の管理＞例えば、ホストが接続する記憶サブシステム1と記憶サブシステム2とが、記憶サブシステム1から記憶サブシステム2にデータを複写する非同期転送で運用されているとする。この場合、もし、記憶サブシステム1におけるデータの書き込み順と、記憶サブシステム2におけるデータの書き込み順とが異なると、両記憶サブシステム1、2におけるデータの整合性が保証されなくなる。以下、このようなデータの不整合を回避するための

仕組みについて説明する。

【0147】まず、各記憶サブシステム1、2における記憶資源の記憶領域に所定サイズ（例えば、16Kバイトごと）のブロックを区画して各ブロックに固有のブロック番号を割り当てる。そして、ホストからデータの書き込みがあったブロックについて、そのブロック番号とデータの書き込み順に付与したシーケンス番号との対応づけを制御メモリ6に管理する。例えば、図12に示すように、ブロック番号が56～59のブロックにデータが書き込まれた場合には、図13に示すデータ管理情報を制御メモリ6に作成する。

【0148】記憶サブシステム1から記憶サブシステム2への非同期転送に際しては、図14の転送データフォーマットに示すように、転送するデータに前記データ管理情報を付帯させる。一方、これを受信した記憶サブシステム2では、図15に示すように、データに付帯して送信されてきた前記データ管理情報を、制御メモリ6に管理する。ここで制御メモリ6に管理される前記データ管理情報、すなわち、シーケンス番号とブロックIDの組み合わせには、これに対応するデータのキャッシュメモリ上の位置情報も対応づけて記憶されている。記憶サブシステム2は、前記データ管理情報のシーケンス番号の順番にこれに対応するキャッシュメモリ上の前記位置情報に記憶されているデータを記憶資源に書き込んでいく。

【0149】以上のようにして、ホストが記憶サブシステム1の記憶資源に書き込んだ順番どおりに、記憶サブシステム2の記憶資源においてもデータが書き込まれ、両記憶サブシステム1、2におけるデータの整合が保証されることになる。

【0150】＜マルチホップ方式＞図16（a）に示す広域データストレージシステムは、サイト1に設置された記憶サブシステム1と、サイト2に設置された記憶サブシステム2と、サイト3に設置された記憶サブシステム3とを備える。記憶サブシステム1には、この記憶サブシステム1を記憶手段として利用するホストが接続する。記憶サブシステム1と記憶サブシステム3との間も通信手段により接続される。

【0151】記憶サブシステム1と記憶サブシステム2とは、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送で運用されている。また、記憶サブシステム2と記憶サブシステム3とは、記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチホップ方式」と称する。なお、マルチホップ方式における各記憶サブシステム間の通信を同期転送とするか、非同期転送とするかは任意に設定される。また、これら以外の転送方式であってもよい。

【0152】つぎに、図16（b）とともにマルチホッ

プ方式によるデータ差分管理の詳細について説明する。

【0153】記憶サブシステム1は、ホストから書き込み対象データとその書き込み要求（Write I/O）とを受信すると（S121）、書き込み対象データを自身の論理ボリューム（第1の記憶資源）に書き込むとともに、書き込み処理を行った順にシーケンス番号を付与し、これと前記データが書き込まれた論理ボリューム（第1の記憶資源）上の位置（格納位置）を特定する書き込み位置情報とを対応づけて（所定のテーブルに）記憶する（S122）。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【0154】つぎに、記憶サブシステム1は、前記書き込み対象データを、これに付与された前記シーケンス番号とともに記憶サブシステム2に送信する（S123）。ここでこのような記憶サブシステム間で行われる、データとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには必要に応じて、前述したデータの書き込み位置情報が付帯される。

【0155】記憶サブシステム2は、記憶サブシステム1から送られてくる前記書き込み対象データとシーケンス番号とを受信して、これを自身の論理ボリューム（第2の記憶資源）に書き込む。記憶サブシステム2は、前記書き込み処理が完了すると、その完了通知を記憶サブシステム1に送信する。

【0156】記憶サブシステム2は、記憶サブシステム3に対し、適宜なタイミングで前記書き込み対象データと前記シーケンス番号とを送信する（S124）（なお、図16（b）では、時間差を表現するため、記憶サブシステム1から記憶サブシステム2に送信されるデータのシーケンス番号と、記憶サブシステム2から記憶サブシステム3に送信されるデータのシーケンス番号とを変えて表記している）。

【0157】つぎに、記憶サブシステム3は、前記データと前記シーケンス番号とを受信すると、前記書き込み対象データに対応して発行した前記シーケンス番号を、記憶サブシステム1に送信する（S125）。記憶サブシステム1は、記憶サブシステム3から送られてくるシーケンス番号を受信する。

【0158】ここで記憶サブシステム1は、受信したシーケンス番号と、自身が記憶しているシーケンス番号とこれに対応する書き込み位置情報との対応づけ（テーブル）を対照することで、記憶サブシステム3の論理ボリューム（第3の記憶資源）に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム3から受領した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することにより行われる（S126）。

【0159】以上のようにしてマルチホップ方式におけ

る通常運用が行われる。

【0160】 つぎに、災害等により記憶サブシステム2が停止した場合の回復処理について説明する。

【0161】 図17(a)に示すように、記憶サブシステム1は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム2の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが途切れるなどして、記憶サブシステム1が記憶サブシステム2の障害発生を検知した場合に、記憶サブシステム1と記憶サブシステム3の間を、差分データのみを複写することによってその内容を一致させ、その後記憶サブシステム1と記憶サブシステム3の間を、非同期転送での臨時運用へ移行させる処理について、図17(b)とともに説明する。

【0162】 記憶サブシステム1は、記憶サブシステム2の障害発生を検知した場合(S131)、まず、制御メモリ6上に、自身の論理ボリューム(第1の記憶資源)の所定ブロック単位の前記データ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについての前記シーケンス番号と前記書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップに対応する位置のビットをオンにする(S132)。

【0163】 つぎに、記憶サブシステム1の論理ボリュームの、前記ビットマップ上のオンになっている位置に格納されている差分データを、記憶サブシステム1から記憶サブシステム3の対応する格納位置に複写する(S133)。そして、この複写完了後、記憶サブシステム1から非同期転送により差分データが複写される形態で、臨時運用を開始される(S134)。

【0164】 ここでこの臨時運用への切り替えに際しては、記憶サブシステム2に障害が発生した場合でも、記憶サブシステム1のデータを記憶サブシステム3に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば、記憶サブシステム1と記憶サブシステム3との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを容易に同期させることができる。

【0165】 つぎに、記憶サブシステム2が復旧し、臨時運用から通常運用に切り替える際の一連の処理について説明する。

【0166】 まず、記憶サブシステム1は、自身の論理ボリューム(第1の記憶資源)に記憶している全てのデータを記憶サブシステム2の論理ボリューム(第2の記憶資源)に複写した後、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送での運用を開始する。すなわち、記憶サブシステム1は、ホストからの指示により自身の論理ボリューム(第1の記憶資源)にデータ書き込みを行った場合、書き込んだデータとシ

ーケンス番号とを記憶サブシステム2に送信する。

【0167】 記憶サブシステム2は、記憶サブシステム1から送られてくる前記書き込んだデータとシーケンス番号とを受信して、これを自身の論理ボリューム(第2の記憶資源)に書き込む。記憶サブシステム2は、前記書き込み処理が完了すると、自身の論理ボリューム(第2の記憶資源)に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号とを対応づけて(所定のテーブルに)記憶する。この段階のデータ転送状態を図18に示す。

【0168】 つぎに、記憶サブシステム3は、記憶サブシステム1から送られてくる前記データと前記シーケンス番号とを受信して、前記データを自身の論理ボリューム(第3の記憶資源)に記憶するとともに(図18)前記対応づけにおける前記シーケンス番号を記憶サブシステム2に送信する(図示せず)。

【0169】 記憶サブシステム2は、記憶サブシステム3から送られてくるシーケンス番号を受信する。ここで記憶サブシステム2は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号と、これに対応する書き込み位置情報とを対照することで、記憶サブシステム3の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【0170】 つぎに、臨時運用において記憶サブシステム1から記憶サブシステム3に複写する非同期転送の運用を停止する。この停止後、記憶サブシステム2は、自身の制御メモリ上に、自身の論理ボリューム(第2の記憶資源)の所定ブロック単位の前記データ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについての前記シーケンス番号と書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【0171】 つぎに、記憶サブシステム2は、前記ビットマップにより把握した、記憶サブシステム3の論理ボリューム(第3の記憶資源)において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム3に送信する。

【0172】 記憶サブシステム3は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを、自身の論理ボリューム(第3の記憶資源)の、前記書き込み位置情報により指定される該当データの格納位置に記憶する。これにより、記憶サブシステム2の論理ボリューム(第2の記憶資源)の内容と、記憶サブシステム3の論理ボリューム(第3の記憶資源)の内容との同期が取れることになる。以上の処理終了後、記憶サブシステム2と記憶サブシステム3との間の非同期転送による運用を開始され、図19に示す通常状態での運用が再開する。

【0173】以上のようにして臨時運用から通常運用への切り替えが完了する。

【0174】＜マルチコピー方式＞図20に示す広域データストレージシステムは、サイト1に設置された記憶サブシステム1と、サイト2に設置された記憶サブシステム2と、サイト3に設置された記憶サブシステム3とを備える。記憶サブシステム2にはこの記憶サブシステム2を記憶手段として利用するホストが接続する。なお、記憶サブシステム1と記憶サブシステム3との間も通信手段により接続される。

【0175】記憶サブシステム1と記憶サブシステム2とは、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送で運用されている。また、記憶サブシステム2と記憶サブシステム3とは、記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチコピー方式」と称する。なお、マルチコピー方式において、各記憶サブシステム間の通信を同期転送とするか、非同期転送とするかは前記の形態に限られず、任意に設定される。また、同期転送や非同期転送以外の転送方式であってもよい。

【0176】つぎに、図20とともにこの実施例のデータ差分管理方式について説明する。記憶サブシステム2は、ホストから書き込み対象データとその書き込み要求（Write I/O）を受信すると（S161）、書き込み対象データを自身の論理ボリューム（第2の記憶資源）に書き込む。また、記憶サブシステム2は、書き込まれたデータと、書き込み処理を行った順に付与したシーケンス番号とを、記憶サブシステム1に送信する（S162）。そして同時に、前記書き込まれたデータと前記付与したシーケンス番号とを、記憶サブシステム3に送信する（S164）。なお、前述のマルチホップ方式の場合と同様に、このような記憶サブシステム間で行われるデータとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには、必要に応じて前述したデータの書き込み位置情報が付帯される。

【0177】つぎに、記憶サブシステム1は、記憶サブシステム2から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム（第1の記憶資源）に書き込む。その際、前記シーケンス番号と、これと前記データが書き込まれた論理ボリューム（第1の記憶資源）上の位置（格納位置）を特定する書き込み位置情報とを対応づけて（所定のテーブルに）記憶する（S163）。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【0178】つぎに、記憶サブシステム3は、記憶サブシステム2から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム（第3の記憶資源）に書き込

む。書き込みが完了すると、記憶サブシステム3は記憶サブシステム1に対し、前記書き込み対象データとこれと対になっていた前記シーケンス番号とを記憶サブシステム1に送信する（S165）。記憶サブシステム1は、記憶サブシステム3から送られてくるシーケンス番号を受信する。

【0179】ここで記憶サブシステム1は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号とこれに対応する書き込み位置情報との対応づけを対照することで、記憶サブシステム3の論理ボリューム

（第3の記憶資源）に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム3から受信した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することで行われる（S166）。

【0180】以上のようにしてマルチコピー方式における通常運用が行われる。

【0181】つぎに、災害等により記憶サブシステム2が停止した場合の回復処理について説明する。

【0182】図21（a）に示すように、記憶サブシステム1は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム2の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが途切れるなどして、記憶サブシステム1が記憶サブシステム2の障害発生を検知した場合に、記憶サブシステム2に接続するホストに代えて、記憶サブシステム1と記憶サブシステム3の間を、差分データのみを複写することによってその内容を一致させ、その後記憶サブシステム1と記憶サブシステム3の間を、非同期転送での臨時運用へ移行させる処理について、図21（b）とともに説明する。

【0183】記憶サブシステム1は記憶サブシステム2の障害発生を検知した場合（S171）、例えば、オペレータの操作により、記憶サブシステム2に接続していたホストの業務の運用が、記憶サブシステム1に接続する副ホストに引き継がれる。

【0184】つぎに、記憶サブシステム1は、制御メモリ6上に、自身の論理ボリューム（第1の記憶資源）の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについてのシーケンス番号とデータ更新位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする（S172）。

【0185】つぎに、記憶サブシステム1の論理ボリュームの、前記ビットマップ上のオンになっている位置に対応する位置に格納されている差分データを、記憶サブシステム1から記憶サブシステム3に複写する（S173）。そして、複写完了後、記憶サブシステム1から同期転送によりデータが複写される形態で、臨時運用が開

始される (S174)。

【0186】ここでこの臨時運用への切り替えに際しては、記憶サブシステム2に障害が発生した場合でも、記憶サブシステム1のデータを記憶サブシステム3に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば記憶サブシステム1と記憶サブシステム3との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを簡単に同期させることができる。

【0187】つぎに、記憶サブシステム2が復旧し、臨時運用から通常運用に切り替える際の一連の処理について説明する。

【0188】まず、記憶サブシステム1は、自身の論理ボリューム (第1の記憶資源) に記憶している全てのデータを記憶サブシステム2の論理ボリューム (第2の記憶資源) に複写した後、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送での運用を開始する。なお、このとき記憶サブシステム1と記憶サブシステム3間での非同期転送も継続して行われる。

【0189】記憶サブシステム1は、ホストから書き込まれたデータと、書き込み処理を行った順に付与したシーケンス番号とを、記憶サブシステム2に送信する。そして同時に、前記書き込まれたデータと前記付与したシーケンス番号とを、記憶サブシステム3にも送信する。

【0190】記憶サブシステム2は、自身の論理ボリューム (第2の記憶資源) に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけを記憶する (位置情報管理テーブル作成)。この段階での運用状態を図22に示す。

【0191】記憶サブシステム3は、記憶サブシステム1から送られてくる前期データと前記シーケンス番号とを受信して、前記データを自身の論理ボリューム (第3の記憶資源) に記憶するとともに前記対応づけにおける前記シーケンス番号を記憶サブシステム2に送信する。

【0192】記憶サブシステム2は、記憶サブシステム3から送られてくるシーケンス番号を受信する。ここで記憶サブシステム2は、前記受信したシーケンス番号と、自身が記憶している前記対応づけを対照することで、記憶サブシステム3の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【0193】つぎに、臨時運用において記憶サブシステム1から記憶サブシステム3に複写する非同期転送の運用を停止する。この停止後、記憶サブシステム2は、自身の制御メモリ上に、自身の論理ボリューム (第2の記憶資源) の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについてのシーケンス番号と書き込み位置情報との対応づけに基

づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【0194】つぎに、記憶サブシステム2は、前記ビットマップにより把握した、記憶サブシステム3の論理ボリューム (第3の記憶資源) において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム3に送信する。

【0195】記憶サブシステム3は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて自身の論理ボリューム (第3の記憶資源) に記憶する。これにより、記憶サブシステム2の論理ボリューム (第2の記憶資源) の内容と、記憶サブシステム3の論理ボリューム (第3の記憶資源) の内容との同期が取れることになる。それから記憶サブシステム2から記憶サブシステム3への非同期転送が開始される。この段階での運用状態を図23に示す。

【0196】ここで記憶サブシステム1に接続するホストの記憶サブシステム1へのデータ書き込み処理が完了しており、記憶サブシステム1と記憶サブシステム2の同期が取れている時に、記憶サブシステム1から記憶サブシステム2に対して行っていたデータの複写を、記憶サブシステム2から記憶サブシステム1に対して行うように切り替える。すなわち、同期が取れている状態で切り替えを行うことで、差分データを複写する等の作業が必要でなくなる。

【0197】つぎに、記憶サブシステム1に接続するホストにより運用されている業務を、記憶サブシステム2に接続するホストに引き継ぐ。そして、記憶サブシステム2から記憶サブシステム3にデータを複写する同期転送による運用を開始することで、図24に示す通常状態での運用が再開することになる。

【0198】以上のようにして臨時運用から通常運用への切り替えが完了する。

【0199】<他の障害復旧方式>つぎに、障害復旧方式のバリエーションについて説明する。

【0200】図25に示すマルチホップ方式において、記憶サブシステム1がダウンした場合 (a) には、記憶サブシステム2に副ホストを接続し、この副ホストにより記憶サブシステム1に接続するホストの業務を引き継ぐ。なお、記憶サブシステム2と記憶サブシステム3の間では、非同期転送での運用が行われている (b)。

【0201】記憶サブシステム1が復旧した場合には、まず、記憶サブシステム2の全データを記憶サブシステム1に複写し、副ホストの業務を記憶サブシステム1に接続するホストに引き継ぐ。そして、前記の要領で、記憶サブシステム1と記憶サブシステム2との間のデータ転送方向を逆向きにすることにより、通常運用を再開する (c)。

【0202】図26に示すマルチホップ方式において、

記憶サブシステム3に障害が発生した場合（a）には、記憶サブシステム3の復旧後、記憶サブシステム2から記憶サブシステム3に全データを複写して記憶サブシステム3のデータを記憶サブシステム2と同期させ、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する（b）。

【0203】図27に示すマルチコピー方式において、記憶サブシステム1に障害が発生した場合（a）には、記憶サブシステム1の復旧後、記憶サブシステム2から記憶サブシステム1に全データを複写して記憶サブシステム1のデータを記憶サブシステム2と同期させ、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する（b）。

【0204】図28に示すマルチコピー方式において、記憶サブシステム3に障害が発生した場合には、記憶サブシステム3の復旧後、記憶サブシステム2から記憶サブシステム3に全データを複写して記憶サブシステム3のデータを記憶サブシステム2と同期させ、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する。

【0205】＜複写元・複写先、書き込み位置情報の管理＞記憶サブシステム間でデータを転送する場合、データの転送元や転送先の設定や、その転送が同期・非同期いずれの方式で行われるかといった設定は、オペレータが各記憶サブシステムを操作して設定する場合（なお、この場合には、例えば、ある記憶サブシステムが障害を起して使えなくなった場合に、どの記憶サブシステムが次のデータの転送元になり、どの記憶サブシステムが次の転送先になるのかということを、システムの構成時に予め登録しておく）、記憶サブシステムに付帯するシステムが自動的に行うようにしている場合など、システムの構成に応じて様々な形態で行われる。

【0206】また、シーケンス番号と書き込み位置情報の対応づけの管理は、例えば、オペレータが、転送元や転送先を記憶サブシステムに登録する操作を開始する契機で行う。

【0207】＜記憶サブシステムの選択方式＞図29に示す広域データストレージシステムは、記憶サブシステム1とこれに接続するホスト1h、記憶サブシステム1からデータが非同期転送される記憶サブシステム2および記憶サブシステム3を備えている。ホスト1hもしくは記憶サブシステム1に障害が発生した場合、迅速に記憶サブシステム2もしくは記憶サブシステム3のどちらか一方を主たる記憶サブシステムとして選択し、また、

信頼性・保全性確保のため、これら2つの記憶サブシステム2および3においてデータを2重化管理する。以下、ホスト1hもしくは記憶サブシステム1に障害が発生した場合に行われる処理について説明する。

【0208】記憶サブシステム2は、例えば、記憶サブシステム1から送信されてくるデータの有無や、記憶サブシステム1からあらかじめ設定された時間等に送られてくるハートビートメッセージの監視により、ホスト1hや記憶サブシステムに障害が発生したことを検知する。

【0209】障害を検知した場合、記憶サブシステム2は、迅速に主たる記憶サブシステムを決定し、副ホスト2もしくは副ホスト3による臨時運用に切り替える。主たる記憶サブシステムの選択はつぎのように行われる。まず、障害を検知した記憶サブシステム2は、記憶サブシステム3に、前述したシーケンス番号のうち最新のシーケンス番号の送信を要求するメッセージを送信する。記憶サブシステム3は、前記メッセージを受信すると、自身が記憶している最新のシーケンス番号を記憶サブシステム2に送信する。

【0210】記憶サブシステム2は、記憶サブシステム3から送られてきたシーケンス番号と、自身が記憶している最新のシーケンス番号とを比較して、より最新のシーケンス番号を受信している記憶サブシステムを、主たる記憶サブシステムとして選出し、選出した記憶サブシステムの識別子を選出候補として記憶するとともに、前記識別子を記憶サブシステム3に送信する。記憶サブシステム3は、送信されてきた前記識別子を受信し、これによりどの記憶サブシステムが主たるサブシステムとして選出されたのかを認知する。

【0211】なお、以上の選出処理において、記憶サブシステム間の通信方式の性質などの諸事情により、記憶サブシステム2もしくは記憶サブシステム3が記憶しているシーケンス番号に抜けが存在することがある。そこで、このような場合には、連続しているシーケンス番号のうちで最新のものを、前記の比較に用いる。

【0212】主たる記憶サブシステムが選出されると、つぎに、記憶サブシステム2と記憶サブシステム3とによりデータの二重化管理を行うため、両者が記憶しているデータの内容を一致させる。これは、記憶サブシステム間で全データの複写や差分データの複写により行われる。記憶サブシステム間でデータが一致すると、主たる記憶サブシステムとして選出された記憶サブシステムは、自身に接続している副ホストに、自身が主たる記憶サブシステムとなる旨を送信する。副ホストはこれを受信して代行運用を開始する。また、記憶サブシステム2と記憶サブシステム3との間で、同期転送もしくは非同期転送によるデータの二重化管理が開始される。

【0213】なお、以上の説明では、記憶サブシステム2が記憶サブシステム3から最新のシーケンス番号を取

得して、主たる記憶サブシステムを選出するようにしているが、この処理は記憶サブシステム3が行ってもよい。

【0214】また、記憶サブシステム1乃至記憶サブシステム3の3台構成の記憶サブシステムにおいて、記憶サブシステム1の障害発生時に代行して運用される他の記憶サブシステムを選出する仕組みを一例として説明したが、前述の仕組みは、4台以上の記憶サブシステムで構成される広域データストレージシステムにも適用することができる。

【0215】＜キャッシュメモリ上のデータの管理＞ホストが接続する一次の記憶サブシステムに、この一次の記憶サブシステムのデータのリモートコピー先である1以上の二次の記憶サブシステムが接続する系における、一次の記憶サブシステムのキャッシュメモリ上のデータの管理に関する実施例について説明する。

【0216】前記の系において、一次の記憶サブシステムから二次の記憶サブシステムに複写（リモートコピー）する必要の無いデータについては、一次の記憶サブシステムの記憶資源にデータを書き込んだ後は、そのデータを当該記憶サブシステムのキャッシュメモリ上から消去されてもよいが、二次の記憶サブシステムに複写する場合には、少なくともそのデータを二次の記憶サブシステムに送信するまではキャッシュメモリ上に残しておく必要がある。また、転送先となる二次の記憶サブシステムが複数存在する場合には、通信手段の違いや運用上の差異などにより、通常、二次の記憶サブシステムについての転送が同時に行われるわけではないので、このような場合には、全ての二次の記憶サブシステムに対する転送が終了するまで、データをキャッシュメモリ上に残しておく仕組みが必要である。

【0217】そこで、一次の記憶サブシステムにおいて、キャッシュ上に置かれているデータについて、一次の記憶サブシステムに接続する二次の各記憶サブシステムについての転送が完了しているかどうかを管理するようにする。具体的には、例えば、図30に示すように、キャッシュメモリ上に区画された記憶ブロック（#1，～，#n）ごとに、それぞれの記憶ブロックに格納されているデータについて、二次の各記憶サブシステムへの転送が完了しているかどうかを示すテーブルを、一次の記憶サブシステムにおいて管理するようにする。

【0218】なお、このテーブルにおいて、ビット「0」は転送が完了していることを示し、ビット「1」は転送が完了していないことを示す。ホストからのデータが一次の記憶サブシステムに書き込まれた時に、データが書き込まれた記憶ブロックの転送先となっている二次の記憶サブシステムに対応するビットに「1」がセットされる。ある記憶ブロックの「1」がセットされているビットのうち、データの転送が完了した二次の記憶サブシステムについてのビットは、転送完了後に「0」と

なる。

【0219】そして、全ての二次の記憶サブシステムについてのビットが「0」となった記憶ブロックに格納されているデータについては、キャッシュメモリ上から消去してもよいということになる。

【0220】

【発明の効果】図1、図9及び図10で示した、3つ以上のサイトを有する広域データストレージシステムにおいて、いずれかのサイトに、いつ災害・障害が発生しても、巨視的に見て常時、データの順序性を保証した論理ボリュームを残すことができる。

【0221】直接データ転送に関与していなかった論理ボリューム間、例えば、図7の記憶サブシステム1と記憶サブシステム3との間で、差分データのみコピーすれば、即時に、非同期リモートコピーのペアを生成でき、広域データストレージシステムの運用再開が、即時に、可能となる効果がある。

【0222】本発明では、記憶サブシステムの内部にリモートコピーを実施するための冗長な論理ボリュームを必要としないため、記憶サブシステムのメモリー資源の使用効率上がり、記憶サブシステムのコストパフォーマンスが向上する効果がある。

【図面の簡単な説明】

【図1】本発明に係る広域データストレージシステムの全体構成の一例を示した説明図である。

【図2】記憶サブシステムの一例を示した概念図である。

【図3】図1の構成において、データコピー監視機能を説明するための概念図である。

【図4】本発明を実現するための転送状態／ビットマップの一例を示した図である。

【図5】一般的な同期転送によるコピーの制御の概略を説明するための図である。

【図6】非同期リモートコピーの制御の概略を説明するための図である。

【図7】図9の全体構成において、データセンタ2に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図8】図1の全体構成において、データセンタ1に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図9】本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図10】データセンタを4拠点以上設置した場合の、本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図11】図9の全体構成において、データコピー監視機能を説明するための概念図である。

【図12】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶資源

のデータを管理する単位であるブロックの概念を示す図である。

【図13】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、データ管理情報の概念を示す図である。

【図14】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、転送データのフォーマットの概念を示す図である。

【図15】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶サブシステム2において管理されるデータ管理情報の概念を示す図である。

【図16】(a)はマルチホップ方式の広域データストレージシステムの概念を示す図であり、(b)は(a)に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図17】(a)はマルチホップ方式の広域データストレージシステムの概念を示す図であり、(b)は(a)に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図18】マルチホップ方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図19】マルチホップ方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図20】(a)はマルチコピー方式の広域データストレージシステムの概念を示す図であり、(b)は(a)に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図21】(a)はマルチコピー方式の広域データストレージシステムの概念を示す図であり、(b)は(a)に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図22】マルチコピー方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図23】マルチコピー方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図24】マルチコピー方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図25】(a)～(c)は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図26】(a)、(b)は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図27】(a)、(b)は、マルチコピー方式における障害復旧方式の他のバリエーションを説明する図である。

【図28】(a)、(b)は、マルチコピー方式における障害復旧方式の他のバリエーションを説明する図である。

【図29】障害発生時において、本番業務を代行させる記憶サブシステムの選択方法を説明する、広域データストレージシステムの概念図である。

【図30】本発明の一実施例による、キャッシュメモリ上のデータの管理方法における、二次の各記憶サブシステムへのデータの転送状態を管理するテーブルを示す図である。

【符号の説明】

1 記憶サブシステムの制御装置

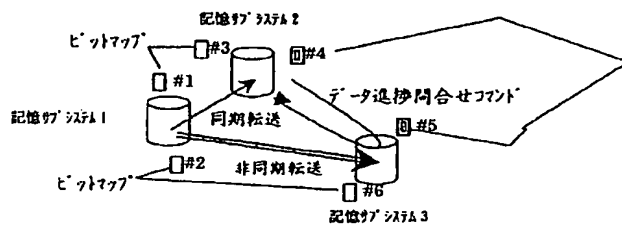
5 キャッシュメモリ

6 制御メモリ

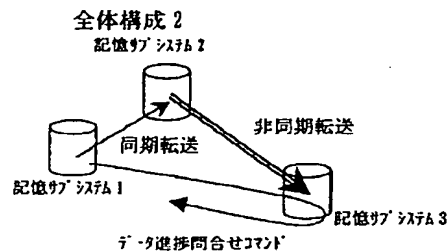
#1～#6 転送状態/ビットマップ。

【図3】

データ監視機能

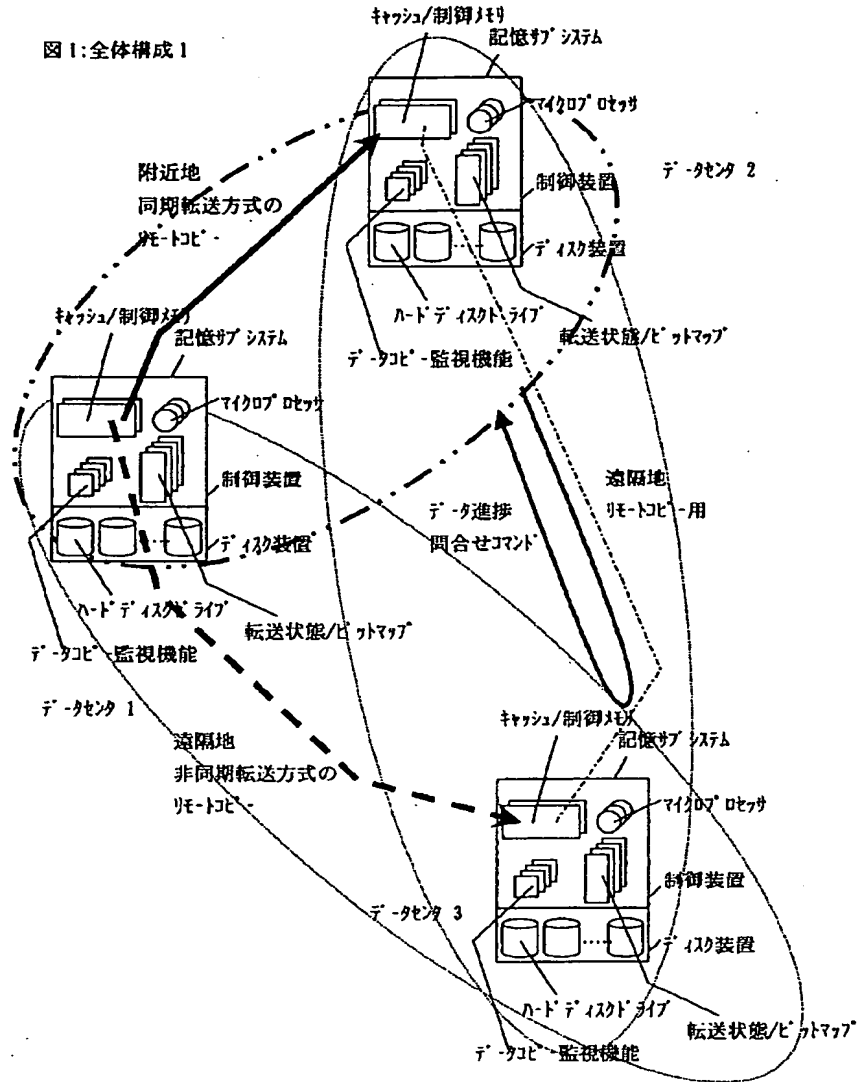


【図9】



【図1】

図1:全体構成1

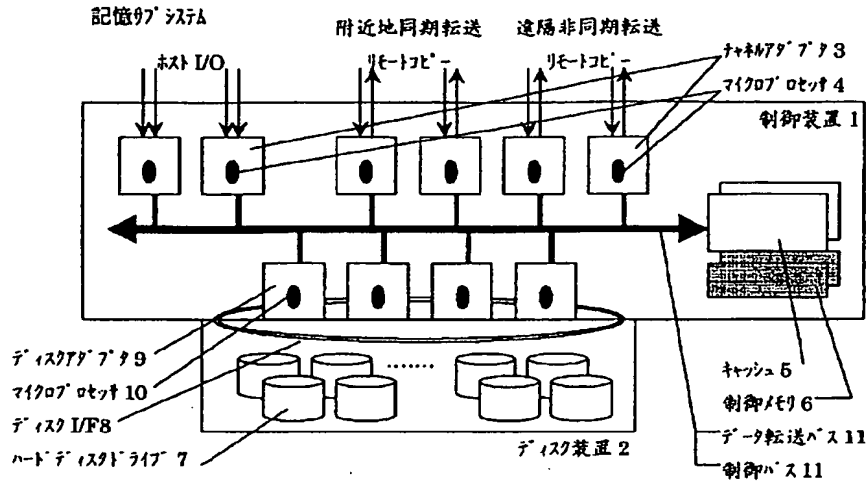


【図4】

転送状態/ビットマップ

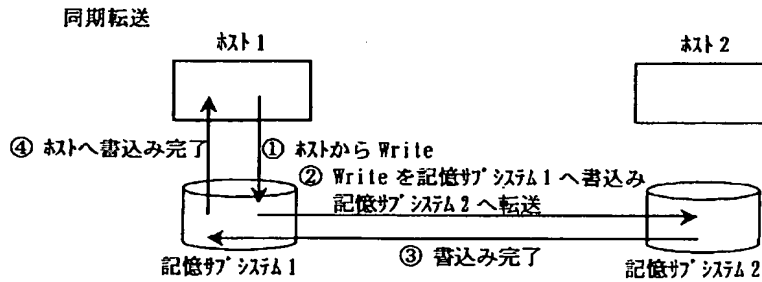
ブロック番号	Update フラグ	カウンタ値
...
3030	1	3	...
3031	0	0	...
3032	1	2	...
...
4032	1	Over Flow	...
...

【図2】



【図5】

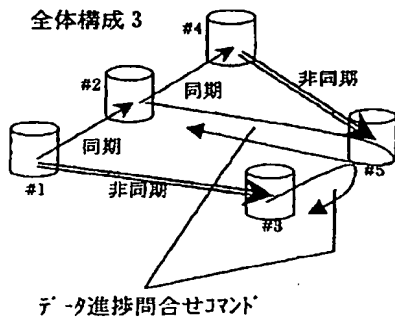
【図13】



【図10】

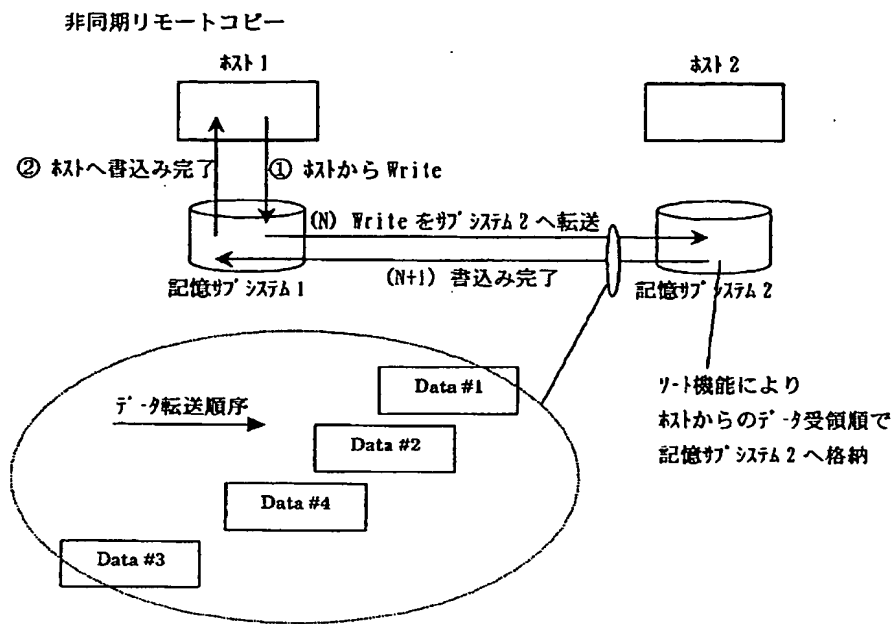
【図15】

シーケンス番号	0
ブロック番号	5 6
シーケンス番号	1
ブロック番号	5 7
シーケンス番号	2
ブロック番号	5 8
シーケンス番号	3
ブロック番号	5 9



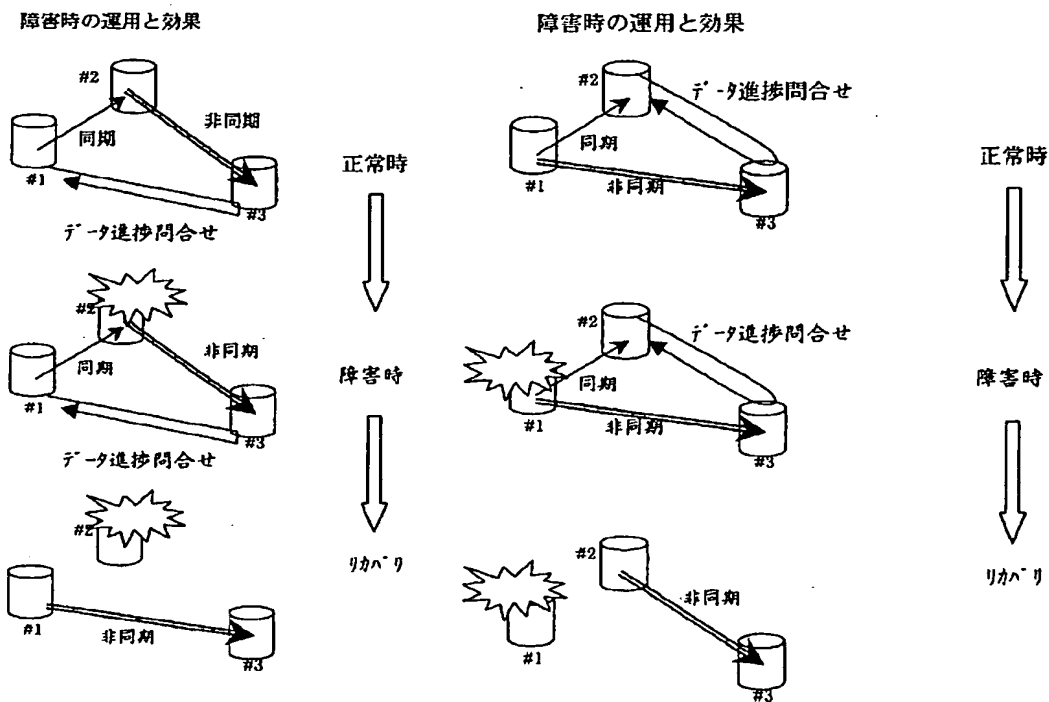
シーケンス番号	0
ブロック番号	5 6
キャッシュ番号	1 2 2
シーケンス番号	3
ブロック番号	5 9
キャッシュ番号	2
シーケンス番号	1
ブロック番号	5 7
キャッシュ番号	1 6

【図 6】

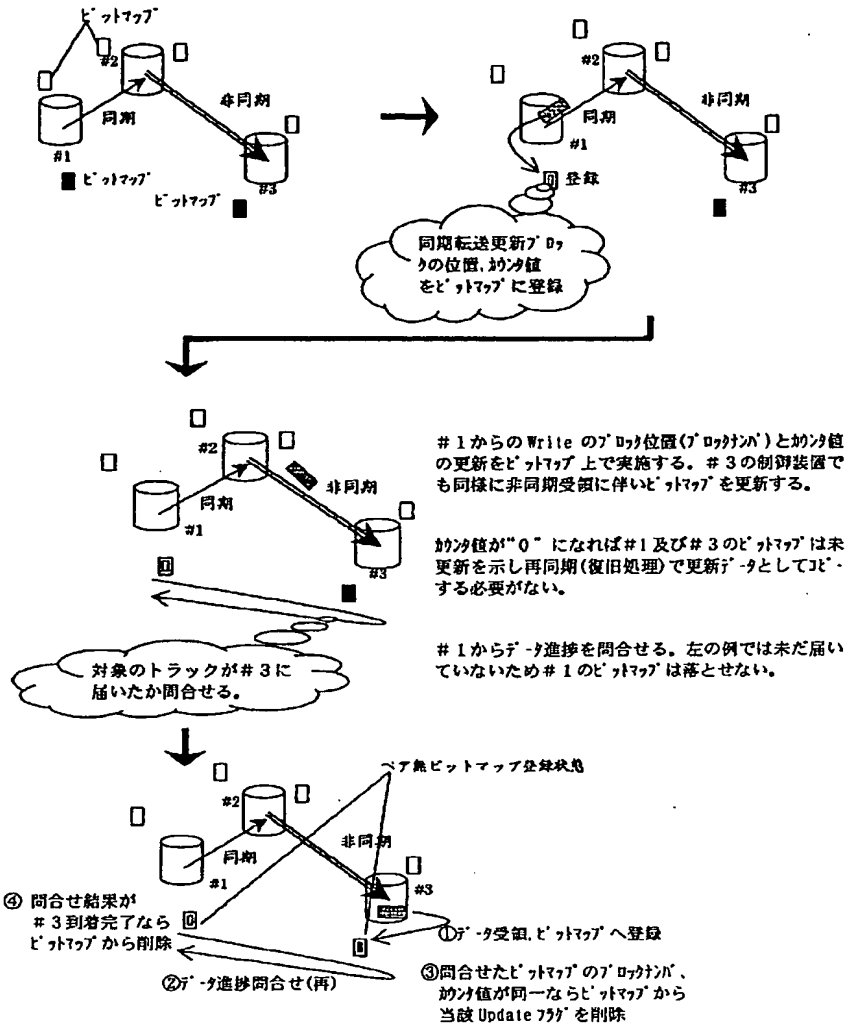


【图7】

【图8】



【図 1 1】



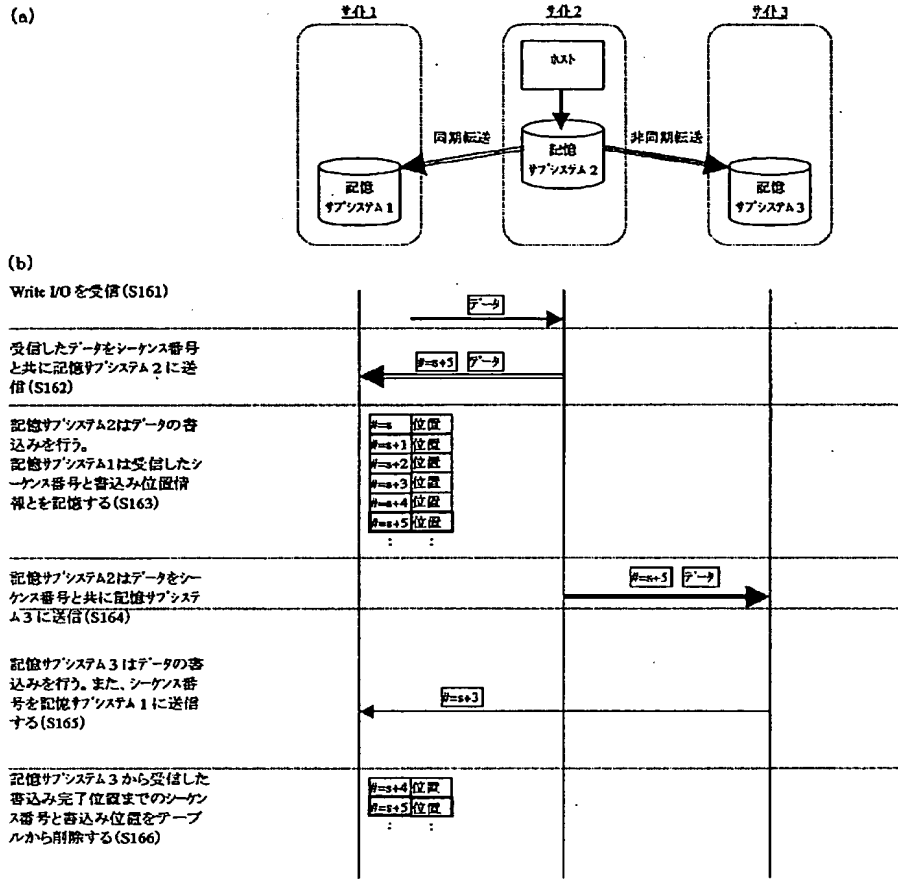
【図 1 2】

ブロック番号56	ブロック番号57	ブロック番号58	ブロック番号59
----------	----------	----------	----------

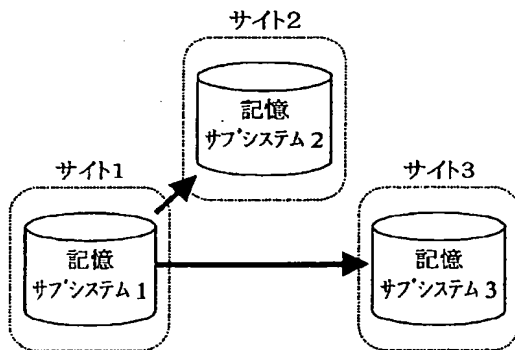
【図 1 4】

ブロック番号 シーケンス番号	データ	ブロック番号 シーケンス番号	データ	ブロック番号 シーケンス番号	データ
-------------------	-----	-------------------	-----	-------------------	-----

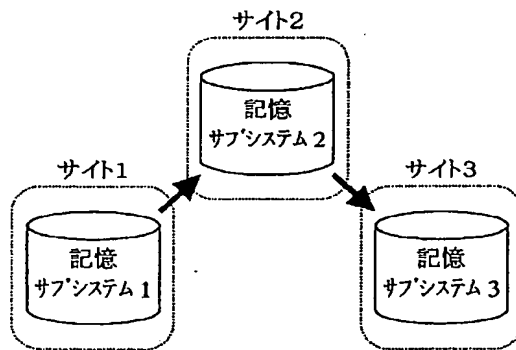
【図 16】



【図 18】

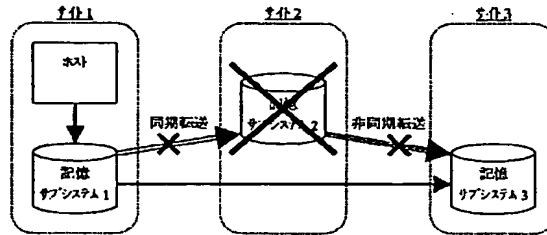


【図 19】



【図 17】

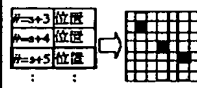
(a)



(b)

記憶サブシステム 2 に障害発生 (S131)

記憶サブシステム 3 との差分を示すビットマップを生成 (S132)



ビットマップに基づいて、記憶サブシステム 1 から記憶サブシステム 3 へ差分データを複写する (S133)

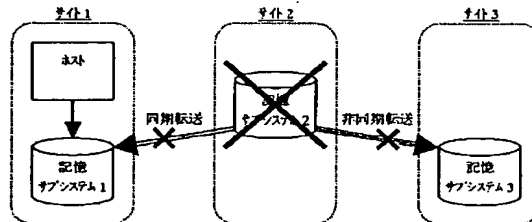


臨時運用開始 (S134)



【図 21】

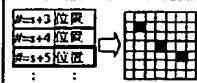
(a)



(b)

記憶サブシステム 2 に障害発生 (S171)

記憶サブシステム 3 との差分を示すビットマップを生成 (S172)



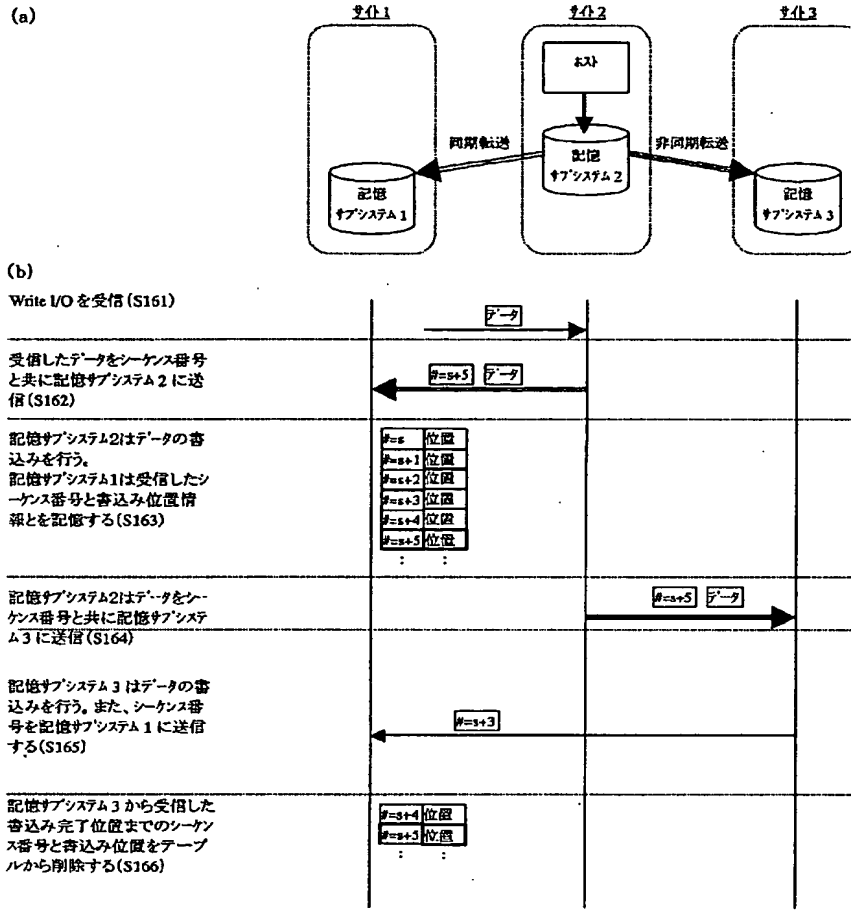
ビットマップに基づいて、記憶サブシステム 1 から記憶サブシステム 3 へ差分データを複写する (S173)。



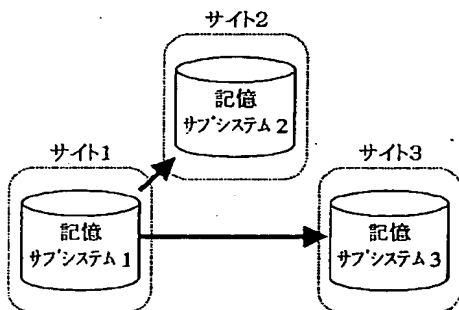
臨時運用開始 (S174)



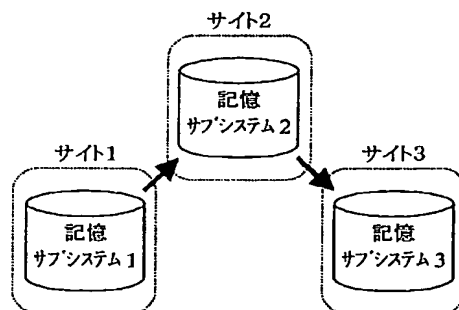
【図20】



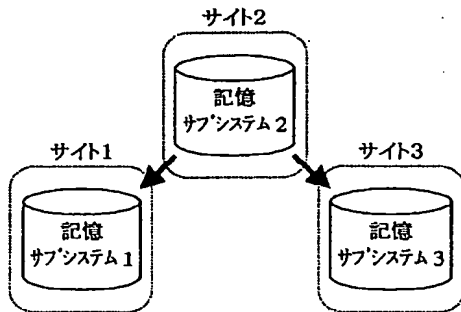
【図22】



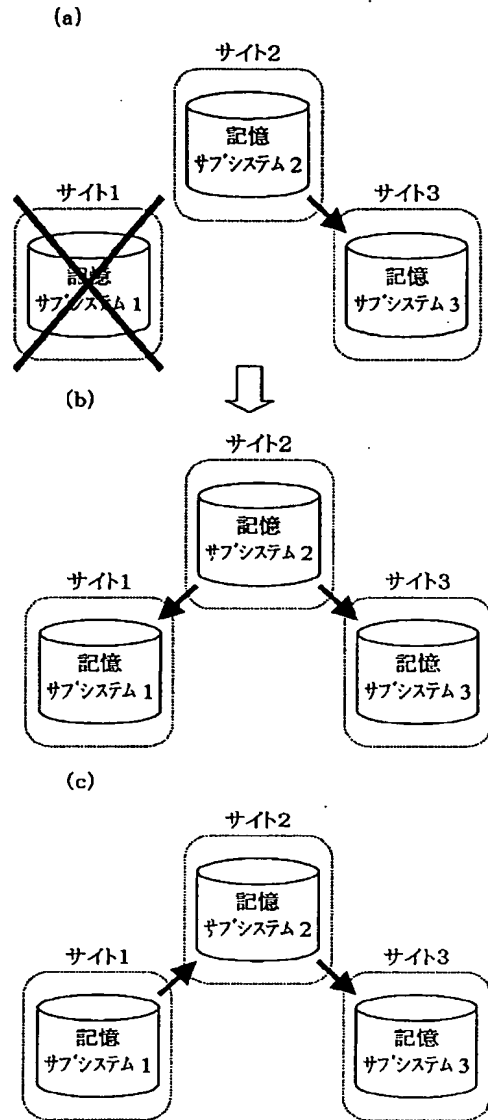
【図23】



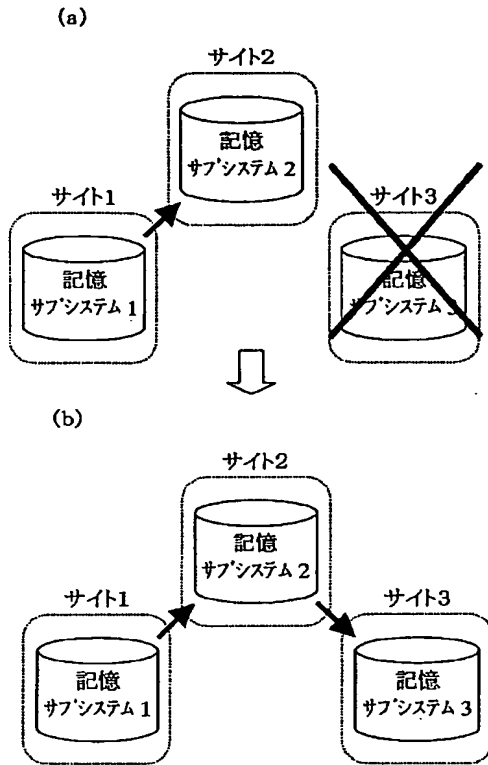
【図24】



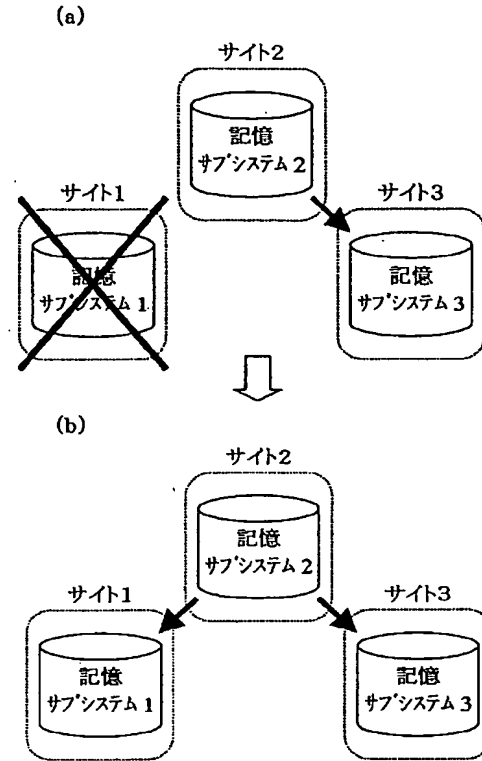
【図25】



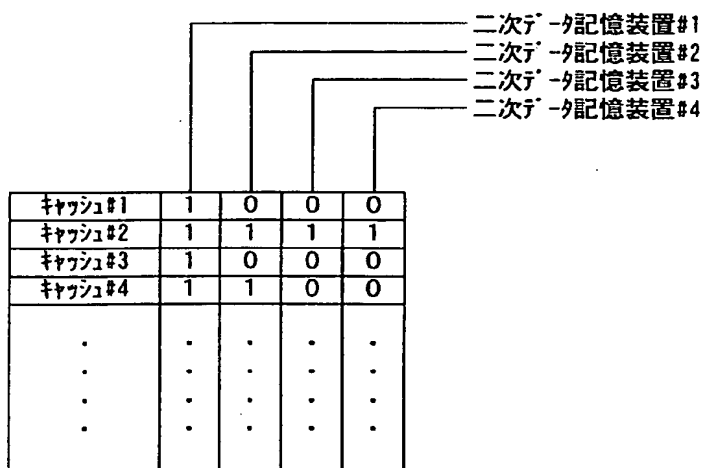
【図26】



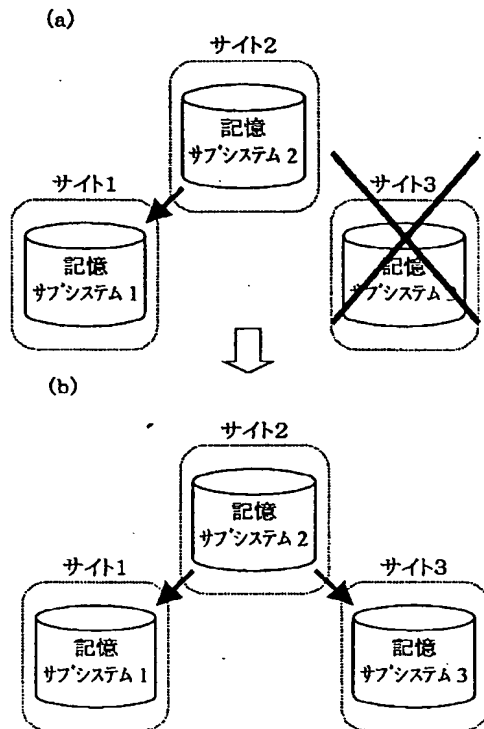
【図27】



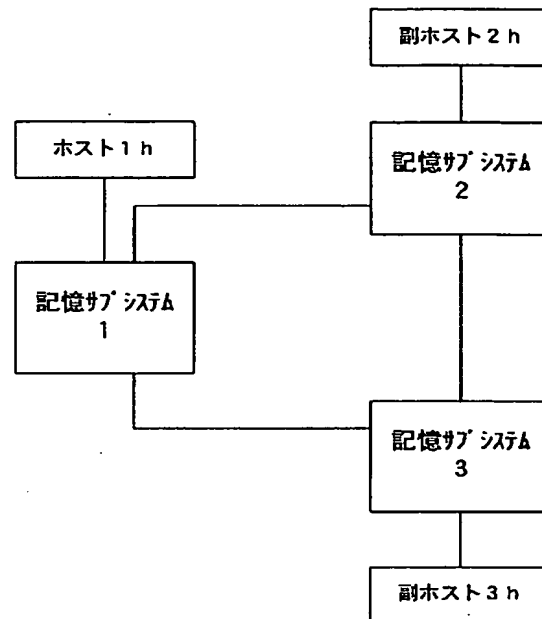
【図30】



【図28】



【図29】



フロントページの続き

(72)発明者 尾形 幹人
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

(72)発明者 岡見 吉規
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

(72)発明者 檜垣 誠一
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

(72)発明者 安部井 大
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

(72)発明者 木城 茂
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内

Fターム(参考) 5B065 BA01 CE01 EA35
5B082 DE03 GA04

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-305856

(43)Date of publication of application : 02.11.2000

(51)Int.Cl.

G06F 12/16

G06F 3/06

G06F 11/20

G06F 12/00

G06F 13/00

(21)Application number : 11-117670

(71)Applicant : HITACHI LTD

(22)Date of filing : 26.04.1999

(72)Inventor : TABUCHI HIDEO

NOZAWA MASASHI

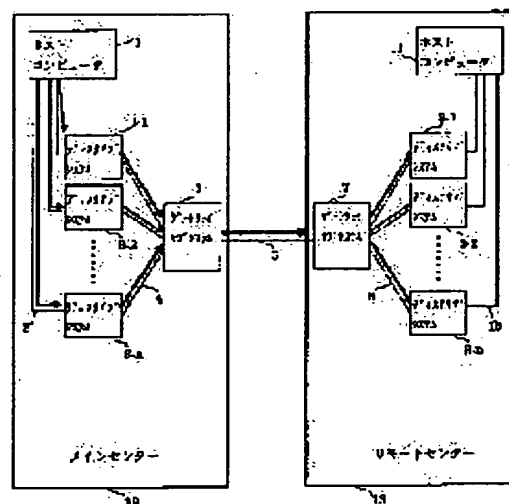
SHIMADA AKINOBU

(54) DISK SUBSYSTEMS AND INTEGRATION SYSTEM FOR THEM

(57)Abstract:

PROBLEM TO BE SOLVED: To guarantee the sequence of update and the consistency of data by doubling data between disk subsystems on a main-center side and a remote-center side through gateway subsystems.

SOLUTION: Data which are written from a host computer 1 are doubled between disk subsystems 3-1, 3-2...3-n and a gateway subsystem 5 and held macroscopically in the same state. The gateway subsystem 5 adds information for holding the sequence of update. Further, the data are doubled between the gateway subsystem 5 and a gateway subsystem 7 by asynchronous remote copying while the sequence of update is guaranteed. The disk subsystems 9-1, 9-2...9-n have the data updated in synchronism with the update of the gateway subsystem 7. Those are all actualized only by the function of the disk subsystems and no new software need not be introduced.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-305856

(P2000-305856A)

(43) 公開日 平成12年11月2日 (2000. 11. 2)

(51) Int.Cl.	識別記号	F I	テマコード (参考)
G 0 6 F 12/16	3 1 0	G 0 6 F 12/16	3 1 0 M 5 B 0 1 8
3/06	3 0 4	3/06	3 0 4 F 5 B 0 3 4
11/20	3 1 0	11/20	3 1 0 C 5 B 0 6 5
12/00	5 3 1	12/00	5 3 1 D 5 B 0 8 2
13/00	3 0 1	13/00	3 0 1 R 5 B 0 8 3
審査請求 未請求 請求項の数9 O L (全 10 頁)			

(21) 出願番号 特願平11-117670

(22) 出願日 平成11年4月26日 (1999. 4. 26)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 田淵 英夫

神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内

(72) 発明者 野沢 正史

神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内

(74) 代理人 100075096

弁理士 作田 康夫

最終頁に続く

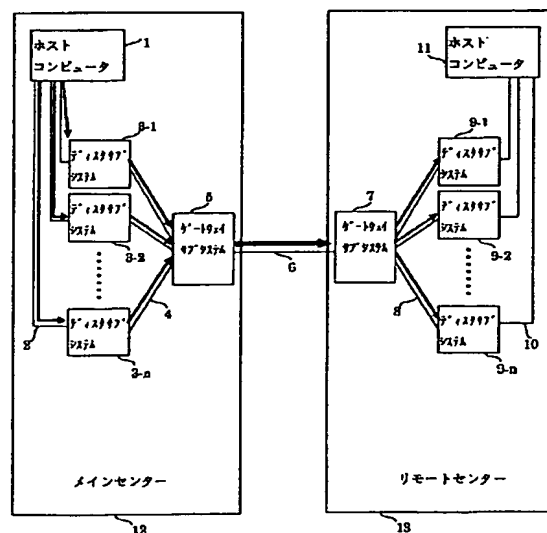
(54) 【発明の名称】 ディスクサブシステム及びこれらの統合システム

(57) 【要約】 (修正有)

【課題】 ホストコンピュータに負担を掛けることなく、複数のディスクサブシステムにわたってデータ更新の順序性／データの整合性を保証でき、導入が容易かつホストコンピュータの性能低下が無い、非同期型のリモートコピー機能を有するディスクサブシステムを提供する。

【解決手段】 各センターのディスクサブシステムのリモートコピーの対象となるボリュームとゲートウェイサブシステムの任意のボリュームの間は同期型リモートコピーでデータの二重化が行なわれ、メインセンターのゲートウェイサブシステムは自サブシステム内のボリュームが更新された順番に従い更新データをリモートセンターのゲートウェイサブシステムに送付しリモートセンターのゲートウェイサブシステムは受け取った順番に従い更新データを自サブシステム内のボリュームに反映する非同期型のリモートコピーでデータの二重化が行なわれるリモートコピーシステム。

図1



【特許請求の範囲】

【請求項1】第1の外部記憶装置と、第2の外部記憶装置とに接続されるディスクサブシステムであって、前記第1の外部記憶装置との間のデータ転送は同期型で、前記第2の外部記憶装置との間のデータ転送は非同期型で、それぞれ行うディスクサブシステム。

【請求項2】上位装置と、

前記上位装置に接続された第1の外部記憶装置と、前記第1の外部記憶装置及び第2の外部記憶装置に接続されたディスクサブシステムを有する統合システムであって、

前記ディスクサブシステムは、前記第1の外部記憶装置との間のデータ転送は同期型で、前記第2の外部記憶装置との間のデータ転送は非同期型で、それぞれ行うことを特徴とする統合システム。

【請求項3】前記第2の外部記憶装置は遠隔地に存在することを特徴とする請求項1記載のディスクサブシステム、又は、

前記第2の外部記憶装置は遠隔地に存在することを特徴とする請求項2記載の統合システム。

【請求項4】前記第2の外部記憶装置との間は、通信回線を介して接続されている請求項1記載のディスクサブシステム、又は、

前記ディスクサブシステムと前記第2の外部記憶装置との間は、通信回線を介して接続されている請求項2記載の統合システム。

【請求項5】前記第1及び前記第2の外部記憶装置が、それぞれ、ディスクサブシステムである請求項1記載のディスクサブシステム、又は、

前記第1及び前記第2の外部記憶装置が、それぞれ、ディスクサブシステムである請求項2記載の統合システム。

【請求項6】上位装置と、前記上位装置に接続された第1の外部記憶装置と、

前記第1の外部記憶装置及び第2の外部記憶装置に接続されたディスクサブシステムを有するメインセンターと、

第3の外部記憶装置及び前記ディスクサブシステムに接続された前記第2の外部記憶装置を有するリモートセンターからなる統合システムであって、

前記上位装置からのデータの更新順序が、リモートセンターにおける前記第3の外部記憶装置に対するデータの更新順序となることを特徴とする統合システム。

【請求項7】前記第6記載の統合システムにおいて、前記第1の外部記憶装置と前記ディスクサブシステムとの間のデータ転送は同期型で、

前記第2の外部記憶装置と前記ディスクサブシステムとの間のデータ転送は非同期型で、それぞれ行う統合システム。

【請求項8】前記非同期型のデータ転送に際し、データに順序に関する情報が付加されている請求項1記載のディスクサブシステム、又は、

前記非同期型のデータ転送に際し、データに順序に関する情報が付加されている請求項2記載の統合システム。

【請求項9】前記データに付加された順序に関する情報がシリアル番号である請求項8記載のディスクサブシステム、又は、

前記データに付加された順序に関する情報がシリアル番号である請求項8記載の統合システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はコンピュータシステムのデータを格納する外部記憶装置及びこれらの統合システムに関し、特に、遠隔地に存在する複数の外部記憶装置群（ディスクサブシステム）と、他の複数の外部記憶装置群とを相互に接続し、上位装置たるホストコンピュータを経由せずに、遠隔地に存在する外部記憶装置（ディスクサブシステム）との間で、データを二重化するリモートコピー技術に関する。ここで、ディスクサブシステムとは、上位装置に対し情報の授受を行う制御部と、情報の格納を行うディスク装置を内蔵する記憶装置をいうものとする。

【0002】

【従来の技術】メインセンターとリモートセンターにそれぞれ設置されているディスクサブシステムの間で、データを二重化して保有する、いわゆる、リモートコピー機能を採用した外部記憶システムが、既にいくつか実用化されている。

【0003】かかる従来技術では、ホストコンピュータが介在してリモートコピーの機能を達成するため、種々の課題があった。

【0004】「同期型と非同期型について」リモートコピー機能は、同期型と非同期型の2種類に大別される。

【0005】同期型とはメインセンター内のホストコンピュータ（上位装置）からディスクサブシステムに、データの更新（書込み）指示が有った場合、その指示対象がリモートコピー機能の対象でもあるときは、そのリモートコピー機能の対象であるリモートセンターにおけるディスクサブシステムに対して、指示された更新（書込み）が終了してから、メインセンターのホストコンピュータに更新処理の完了を報告する処理手順をいう。メインセンターとリモートセンターとの地理的距離に応じて、この間に介在するデータ伝送線路の能力の影響を受け、時間遅れ（伝送時間等）が発生する。同期型は、伝送時間を考慮すると遠隔地といっても現実的には数十Kmが限界であった。

【0006】同期型では、メインセンターとリモートセンターのディスクサブシステムのデータの内容が巨視的にみて常に一致している。このため、メインセンターが

災害等により機能を失った場合であっても、リモートセンター側のディスクサブシステムに災害直前までの状態が完全に保存されているので、リモートセンター側で迅速に処理を再開できる効果がある。尚、巨視的にみて常に一致とは、同期型の機能を実施中には、磁気ディスク装置や電子回路の処理時間の単位 (μsec , msec) で、一致していない状態が有り得るが、データ更新処理完了の時点ではデータは必ず同一の状態になっていることを意味している。これは、リモートセンターへの更新データの反映が終了しない限り、メインセンターの更新処理を完了できないためである。このため、特に、メインセンターとリモートセンターの距離が離れており、データ伝送線路が混雑している場合には、メインセンター側のディスクサブシステムのアクセス性能が大幅に劣化する。

【0007】これに対し非同期型とは、メインセンター内のホストコンピュータからディスクサブシステムに、データの更新(書き込み)指示が有った場合、その指示対象がリモートコピー機能の対象であっても、メインセンター内のディスクサブシステムの更新処理が終わり次第、ホストコンピュータに対し更新処理の完了を報告し、リモートセンターのディスクサブシステムにおけるデータの更新(反映)はメインセンターにおける処理とは非同期に実行する処理手順をいう。このためメインセンター内部で必要とされる処理時間でデータ更新が終了するので、リモートセンターへのデータの格納に起因する伝送時間等はかからない。

【0008】非同期型は、リモートセンターのディスクサブシステムの内容が、メインセンター側のそれに対し、常に一致しているわけではない。このため、メインセンターが災害等により機能を失った場合は、リモートセンター側にデータの反映が完了していないデータが消失することとなる。しかし、メインセンター側のディスクサブシステムのアクセス性能を、リモートコピー機能を実施しない場合と同等レベルとすることができる。

【0009】地震等の天災の際のデータのバックアップを考慮すれば、メインセンターとリモートセンターは100km〜数100km程度、分離する必要がある。また、例えば、100Mbit/secから300Mbit/secクラスの高速通信回線をリモートコピー機能のために使用することも可能ではあるが、ディスクサブシステムの顧客に高額の回線使用料を負担させることとなり、経済的でない。

【0010】「順序性保全について」データの伝送時間の課題の他に、複数のディスクサブシステムを有するメインセンターのバックアップをリモートセンターで行おうとするとき、各々のディスクサブシステムが1対1に対応しなければならない(順序性保全)という課題がある。非同期型リモートコピーでは、リモートセンターへの更新データの反映が、メインセンターでの実際の更新処理の発生時点より遅れて処理されることはやむを得な

い。しかし更新の順序はメインセンターと一致していなければならない。

【0011】一般にデータベース等はデータベース本体と各種ログ情報、制御情報から構成されており、それぞれが関連性を持っている。データ更新の際はデータベース本体に加え、これらログ情報、制御情報をも更新し、システムの整合性が保たれている。したがって更新の順序が崩れた場合、更新順序に関連するこれらの情報の整合性も崩れ、最悪の場合には、データベース全体の破壊につながる可能性がある。

【0012】「ホストコンピュータが介在することについて」メインセンター及びリモートセンターに複数のディスクサブシステムが存在する一般的な環境で非同期型のリモートコピーを実現する場合には、ホストコンピュータがディスクサブシステムへデータの更新を指示する場合、タイムスタンプなどの更新順序に関する情報をデータに付加し、これらの情報に基づいて副側のディスクサブシステムの更新データ反映処理が実行されるのが一般的である。例えば、IBM社のXRC(Extended Remote Copy)機能のように、ホストコンピュータが介在してリモートコピー機能を実現している。

【0013】XRC機能の具体的開示は特開平6-290125(米国特許第5446871号)に詳細になされている。XRC機能においては、メインセンター側のホストコンピュータのオペレーティングシステムとディスクサブシステム、リモートセンター側のホストコンピュータのデータムーバソフトウェアとディスクサブシステムの連携により、更新順序情報の発行、送付、これに基づく更新データ反映処理を実現している。

【0014】

【発明が解決しようとする課題】従来技術(XRC機能)により、メインセンター、リモートセンター間の更新順序性を保証しながら非同期型のリモートコピー機能が実現できる。しかし従来技術では、上位ソフトウェアとディスクサブシステムの双方にXRC機能実現の為に仕組が必要であり、且つ、両者が連携しなければならない。専用の新規ソフトウェアの導入が必要なため、ユーザは、ソフトウェアの導入、設定、検査、CPU負荷増加に伴うシステム設計の見直し等の作業が発生する。このため従来の機能の導入のためには所定の期間を要し、費用が発生するという導入障壁があった。

【0015】また、リモートセンターとの通信回線の容量が十分でない場合に非同期型リモートコピー機能を行うと、リモートセンターへ未反映の更新データが増大するという課題があった。

【0016】本発明の目的は、新規ソフトウェアの導入を必要とせず、ディスクサブシステムの機能のみで、更新の順序性やデータの整合性を保証でき、導入が容易かつメインセンターの性能低下が少ない、非同期型のリモ

ートコピー機能を実現することである。

【0017】本発明の別の目的は、非同期型のリモートコピー機能を大容量のデータ格納を行えるディスクサブシステムに適用することで、ディスクサブシステムの顧客に高額の回線使用料を負担させることなく、リモートコピー機能を実現することにある。

【0018】

【課題を解決するための手段】相互に遠隔地に存在するメインセンターとリモートセンターに、それぞれゲートウェイとなるディスクサブシステム（以下、ゲートウェイサブシステム）を一台ずつ配置し、ゲートウェイサブシステムをデータ伝送線路に接続する。そして、両センターのリモートコピーを実施する必要があるディスクサブシステム全てを、センター内のそれぞれのゲートウェイサブシステムに接続する。

【0019】メインセンターのディスクサブシステムのリモートコピーの対象となるボリュームと、メインセンター内のゲートウェイサブシステムの任意のボリュームとの間は、同期型リモート機能により接続し、データの二重化を行う。これによりメインセンター内のリモートコピー対象のディスクサブシステムのボリュームと、メインセンター内のゲートウェイサブシステムのボリュームにおいて、システムの処理時間の遅れ等を無視できれば、同一のデータが保持される。

【0020】メインセンターとリモートセンターのゲートウェイサブシステムの各ボリュームの間は、非同期型のリモートコピーによりデータの二重化を行う。ただし、メインセンターのゲートウェイサブシステムは、自己のサブシステム内のボリュームが更新された順番に従い、更新データをリモートセンターのゲートウェイサブシステムに送付し、リモートセンターのゲートウェイサブシステムは受け取った順番に従い、更新データを自己のサブシステム内のボリュームに反映する。

【0021】リモートセンター内のゲートウェイサブシステムと各ディスクサブシステムの各ボリュームの間は同期型リモートコピーによりデータの二重化を行う。これにより、リモートセンター側のゲートウェイサブシステムのボリュームとリモートコピー対象のディスクサブシステムのボリュームにおいて、巨視的にみて常に同一のデータが保持される。

【0022】なお、ゲートウェイサブシステムでは、リモートコピー対象のボリュームのデータは自己のゲートウェイサブシステム内のバッファメモリに格納される。従って、バッファメモリ以外に、データ格納用のサブシステムのエリアは通常は必ずしも必要としない。但し、利用可能なサブシステムのエリアがあれば、伝送線路を経由したデータ送受の際に、伝送線路の容量に応じて必要となるサブシステムのエリアを利用することはできる。

【0023】以上の構成により、ディスクサブシステム

の機能のみで、メインセンターの複数のディスクサブシステムと、リモートセンターの複数のディスクサブシステムのデータの二重化を更新の順序性を保ちながら実現できる。リモートセンターへの更新データの反映は、メインセンターの各ディスクサブシステムの更新処理とは非同期に実施することができる。これにより高性能で導入の容易な災害バックアップシステムを提供することができる。また、伝送線路の通信容量に応じて、適宜、サブシステムのエリアを用いることができ、顧客の回線使用料負担を軽減できる効果がある。

【0024】

【発明の実施の形態】以下、図面を参照しながら本発明を汎用コンピュータシステムに適用した場合の一例について説明する。

【0025】図1に、汎用コンピュータシステムを装備した複数のデータセンターにおいて、任意の2つのセンター間でデータの二重化を行うために、本発明を適用したときの構成例を示す。

【0026】メインセンター側の一台又は複数台のディスクサブシステムと、リモートセンター側の一台又は複数台のディスクサブシステムは、ホストコンピュータを介さずに、ゲートウェイサブシステムを介して接続され、両センター間でデータの二重化を行うリモートコピーシステムを実現している。

【0027】図1のメインセンター12において、中央処理装置（ホストコンピュータ）1は、インタフェースケーブル2を介して、ディスクサブシステム3-1、3-2、……、3-nに接続されている。ディスクサブシステム3-1、3-2、……、3-nは、ホストコンピュータ1から参照又は更新されるデータを格納する。ゲートウェイサブシステム5は、インタフェースケーブル4を介して、ディスクサブシステム3-1から3-nと接続される。

【0028】ゲートウェイサブシステム5は、ホストコンピュータがディスクサブシステム3-1等にデータの書込み要求を発行すると、これに同期して当該データを自己のサブシステム内のバッファメモリにも書込む。更に、自己のサブシステム内のバッファメモリにデータが書込まれたこととは非同期に、遠隔地に存在するゲートウェイサブシステム7に対し、データの書込み指示を行う。ゲートウェイサブシステム5は、ディスクサブシステム3-1から3-nの台数にかかわらず、必ず一台で構成される。

【0029】ゲートウェイサブシステム7は、リモートセンター13に設置され、インタフェースケーブル6を介して、メインセンター12のゲートウェイサブシステム5と接続されている。なお、インタフェースケーブル6は、一般の通信回線と接続することも可能である。本例ではこの点も含めてインタフェースケーブル6として記述する。ゲートウェイサブシステム7は、ゲートウェイ

イサブシステム5から受け取ったデータを、書き込み指示のあった順に、自己のサブシステム内のバッファメモリに格納する。ゲートウェイサブシステム7は必ず一台で構成される。

【0030】ディスクサブシステム9-1、9-2、……9-nは、インタフェースケーブル8を介して、ゲートウェイサブシステム7と接続される。ディスクサブシステム9-1等は、メインセンター12からゲートウェイサブシステム7にデータの書き込み要求があった場合には、これに同期して当該データを自己のサブシステム内にも書き込む。

【0031】つまり、ホストコンピュータ1から一台または複数台のディスクサブシステム3-1から3-nに対しデータの書き込み指示があった場合には、リモートセンター13内の一台または複数台のディスクサブシステム9-1から9-nにも同じデータが格納される。図1の矢印は、ホストコンピュータ1から書き込み指示のあったデータの流れを示している。

【0032】ホストコンピュータ11は、リモートセンター13においてディスクサブシステム9-1から9-nとインタフェースケーブル10によって接続され、ディスクサブシステム9-1等に対し、参照及び更新を行う中央処理装置である。ホストコンピュータ11は、メインセンター12のホストコンピュータ1が災害や故障等により本来の機能を果たせなくなった場合に、ホストコンピュータ1の代替となって処理を行うことが出来る。このほか、ディスクサブシステム9-1等に格納されているデータを使用して、メインセンター12のホストコンピュータ1とは異なる処理を、ホストコンピュータ1とは別個独立に実行することができるものである。但し、ホストコンピュータ11がディスクサブシステム9-1等に対し処理を行わない場合には、ホストコンピュータ11は不要である。

【0033】本発明の実施の形態として、データの二重化方法と運用の概略を図2、図3を用いて説明する。

【0034】二重化の対象となるデータが格納されたボリュームやデータセット、ディスクサブシステムは、事前に運用者が選択する。そして、対象ボリュームや対象データセット及びディスクサブシステムと、選択したデータの複製を格納するボリュームやデータセット及びディスクサブシステムとの関係を、予め運用者がホストコンピュータからディスクサブシステムに対し設定しておく。

【0035】上記の選択、設定に際し、専用のコンソールやサービスプロセッサを接続又は装備できるディスクサブシステムの場合には、ホストコンピュータを利用せず、そのコンソールやサービスプロセッサを通じて設定できる。図2のフローはホストから選択・設定を行う場合を示している。

【0036】設定方法としては、上記のボリュームやデ

ィスクサブシステムを意味する具体的なアドレスを指定する方法や、ディスクサブシステム内の制御プログラムによって、アドレスの任意の範囲から選択する方法をとることもできる。初期設定として、パス設定やベア設定を行う例を示してある(図2、201)。

【0037】ホストコンピュータ1(図1)から、ディスクサブシステム3-1、3-2、……、3-n(211)に対し、データの書き込み要求(以下、ライトコマンド)が発行される(図2、202)と、ディスクサブシステム3-1、3-2、……、3-nはライトコマンドにもとづき自己のディスクサブシステム内へデータ格納処理を実行しつつ、ゲートウェイサブシステム5(212)に対し、そのデータのライトコマンドを発行する(203)。ここで、ライトコマンドとは、データを書き込むための指示と書き込みデータそのものとを転送するコマンドである。

【0038】ゲートウェイサブシステム5は、ライトコマンドを受領するとライトコマンドに対する処理を実行する(204)。自己のゲートウェイサブシステム内のバッファメモリへのライトコマンドに対するデータ格納処理が完了すると、ゲートウェイサブシステム5は処理の完了をディスクサブシステム3-1、3-2、……、3-n(211)に報告する。これに伴い、処理が完了した順にライトコマンド番号をライトコマンド毎に付与しておき(205)、自己のサブシステムの処理能力に基づいて決定された契機で、ライトコマンド番号が付与されたライトコマンドを、ライトコマンド番号順にゲートウェイサブシステム7(213)に対し発行する(206)。

【0039】ディスクサブシステム3-1、3-2、……、3-n(211)は、ホストコンピュータ1より発行されたライトコマンドに対し、そのライトコマンドに対する処理、即ち、自己のサブシステム内へのデータ格納処理が完了し、かつ、ゲートウェイサブシステム5(212)から書き込み処理の完了が報告されていること(221)を条件に、ホストコンピュータ1に対しライトコマンドに対する処理の完了報告(222)を行う。

【0040】ゲートウェイサブシステム7(213)は、ゲートウェイサブシステム5(212)から発行されたライトコマンドに付与されているライトコマンド番号により、付与された番号順にライトコマンドを受領していることを確認すると、ライトコマンドに対する処理、即ち、自己のサブシステム内のバッファメモリへのデータ格納処理(301)を行う。これに伴い、ディスクサブシステム9(311)に対し、そのデータのライトコマンドを発行する(302)。ディスクサブシステム9(311)は、ゲートウェイサブシステム7から発行されたライトコマンドを受領すると、ライトコマンドに対する処理、即ち、自己のサブシステム内へのデータ格納処理を実行する(303)。

【0041】ディスクサブシステム9-1、9-2、…
……、9-n (311)は、ライトコマンドに対する処理、即ち、自己のサブシステム内のバッファメモリへのデータ格納処理が完了すると、ゲートウェイサブシステム7に対し、処理の完了報告(321)を行う。ゲートウェイサブシステム7(213)は、自己のサブシステム内へのデータ格納処理が完了し、かつ、ディスクサブシステム9-1、9-2、……、9-nから書き込み処理の完了が報告されていることを条件に、ゲートウェイサブシステム5に対し、ライトコマンドに対する処理完了報告(322)を行う。

【0042】本発明により、ホストコンピュータ1から書込まれたデータは、ディスクサブシステム3-1、3-2、……、3-nと、ゲートウェイサブシステム5の間で二重化され、巨視的にみて常に同一の状態に保たれる。この際にゲートウェイサブシステム5において更新の順序を保持するための情報(通番)が付加される。

【0043】また、ゲートウェイサブシステム5とゲートウェイサブシステム7との間は更新の順序を保証しながら非同期のリモートコピーでデータの二重化が行なわれる。ディスクサブシステム9-1、9-2、……、9-nはゲートウェイサブシステム7の更新に同期してデータが更新される。これらはすべてゲートウェイ機能を有するディスクサブシステムを含めて、ディスクサブシステムの機能のみで実現され、ホストコンピュータの処理能力に対し負担とならない。

【0044】図4に、各ゲートウェイサブシステム内のバッファ領域を用いて、伝送線の通信容量が十分でない場合の動作を説明する。同一の符号は既に説明済みであることを示す。このシステムでは、書き込み要求のあったデータを一時的に格納するバッファ領域を、各ゲートウェイサブシステム内に設けておく。通常の伝送線におけるバッファメモリが溢れることを防止するためである。かかるバッファ領域に格納されたサブシステム

のデータは、伝送線を介してメインセンターからリモートセンターへ送られ、そこでリモートセンター側のバッファ領域を介して、ゲートウェイサブシステムへ入力される。こうすれば、二重化の時間的一致度は減少するものの、大容量の通信回線を使用せずとも、非同期型リモートコピー機能を実現できる。

【0045】

【発明の効果】新規ソフトウェアの導入を必要とせずディスクサブシステムの機能のみで、更新の順序性やデータの整合性を保証でき、導入が容易でかつメインセンターの処理性能の低下が無い非同期型のリモートコピーシステムを実現できる。

【0046】また、伝送線の通信容量に応じて、適宜、サブシステムのエリアを用いることができ、顧客の回線使用料負担を軽減できる効果がある。

【図面の簡単な説明】

【図1】本発明の一実施の形態におけるリモートコピーシステムの全体構成を示す図である。

【図2】リモートコピーシステムの処理の詳細なフローチャートを示す図である。

【図3】図2の続きである、リモートコピーシステムの処理の詳細なフローチャートを示す図である。

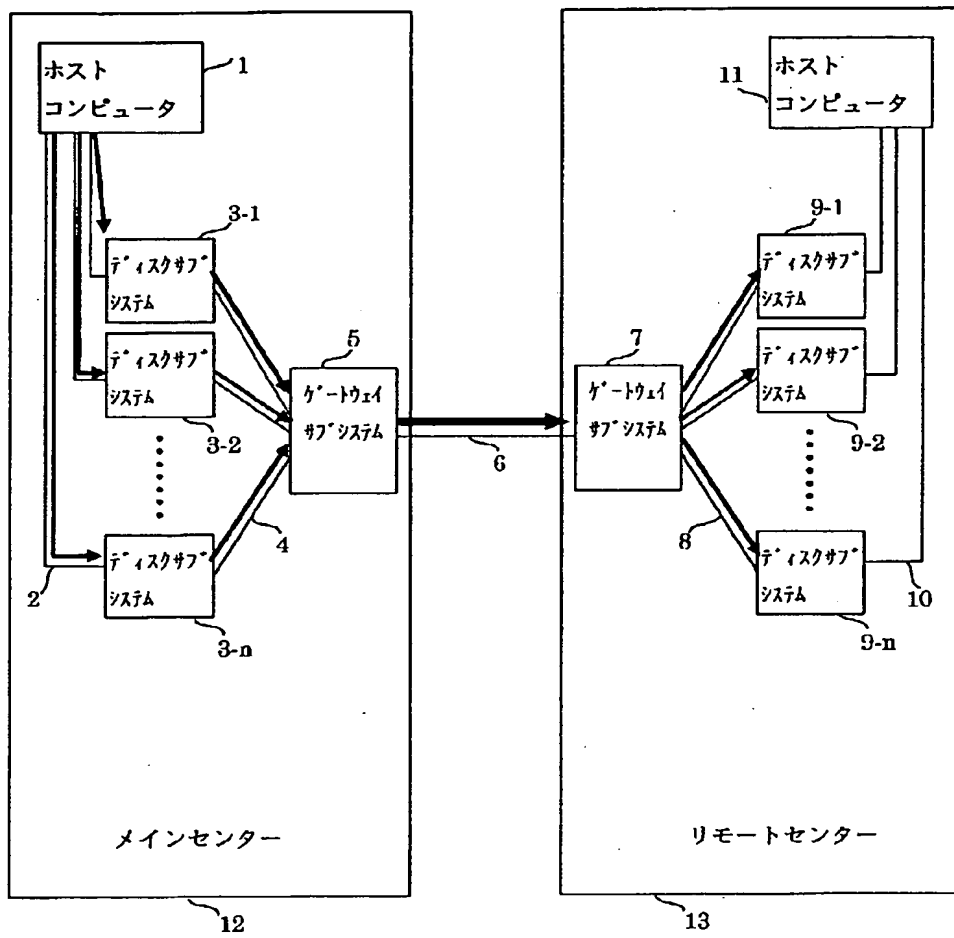
【図4】ゲートウェイサブシステム内にバッファ領域を設けた場合のリモートコピーシステムの処理のフローチャートを示す図である。

【符号の説明】

1…ホストコンピュータ、 2…インタフェースケーブル、 3…ディスクサブシステム、 4…インタフェースケーブル、 5…ゲートウェイディスクサブシステム、 6…インタフェースケーブル、 7…ゲートウェイディスクサブシステム、 8…インタフェースケーブル、 9…ディスクサブシステム、 10…インタフェースケーブル、 11…ホストコンピュータ、 12…メインセンター、 13…リモートセンター。

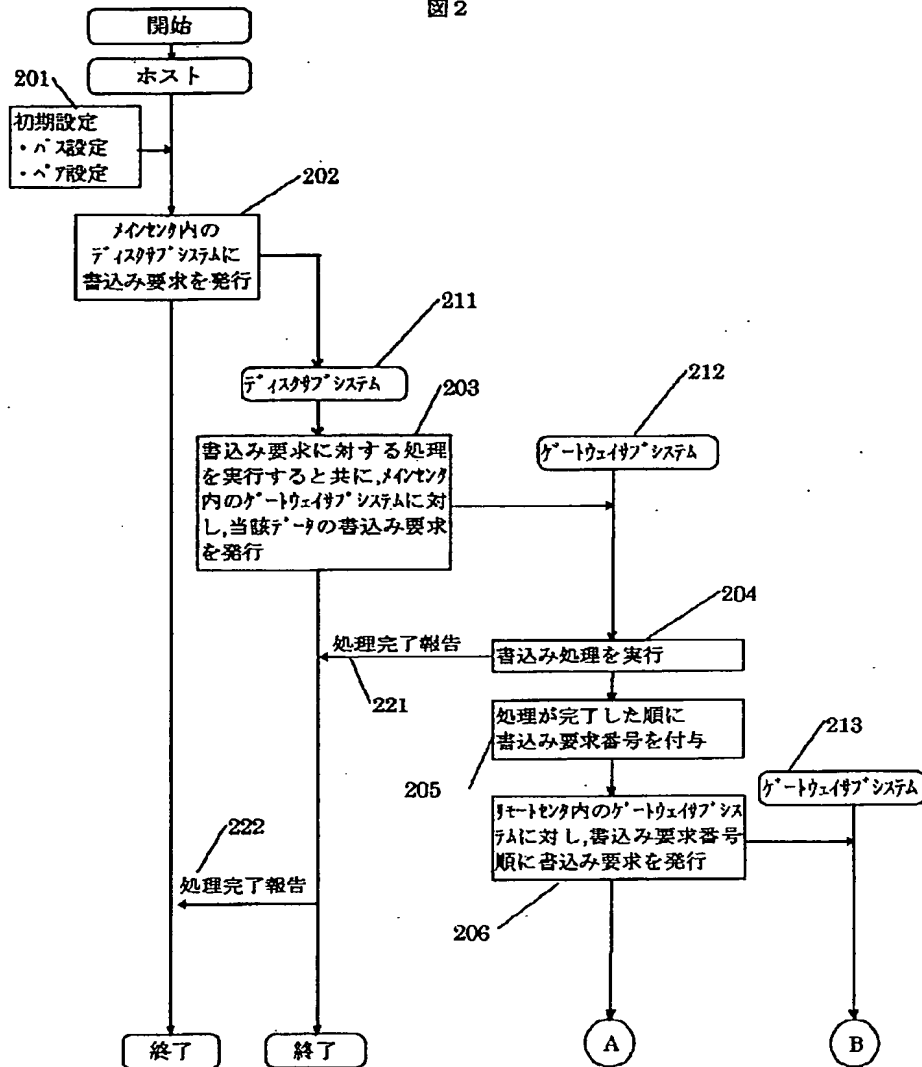
【図1】

図1



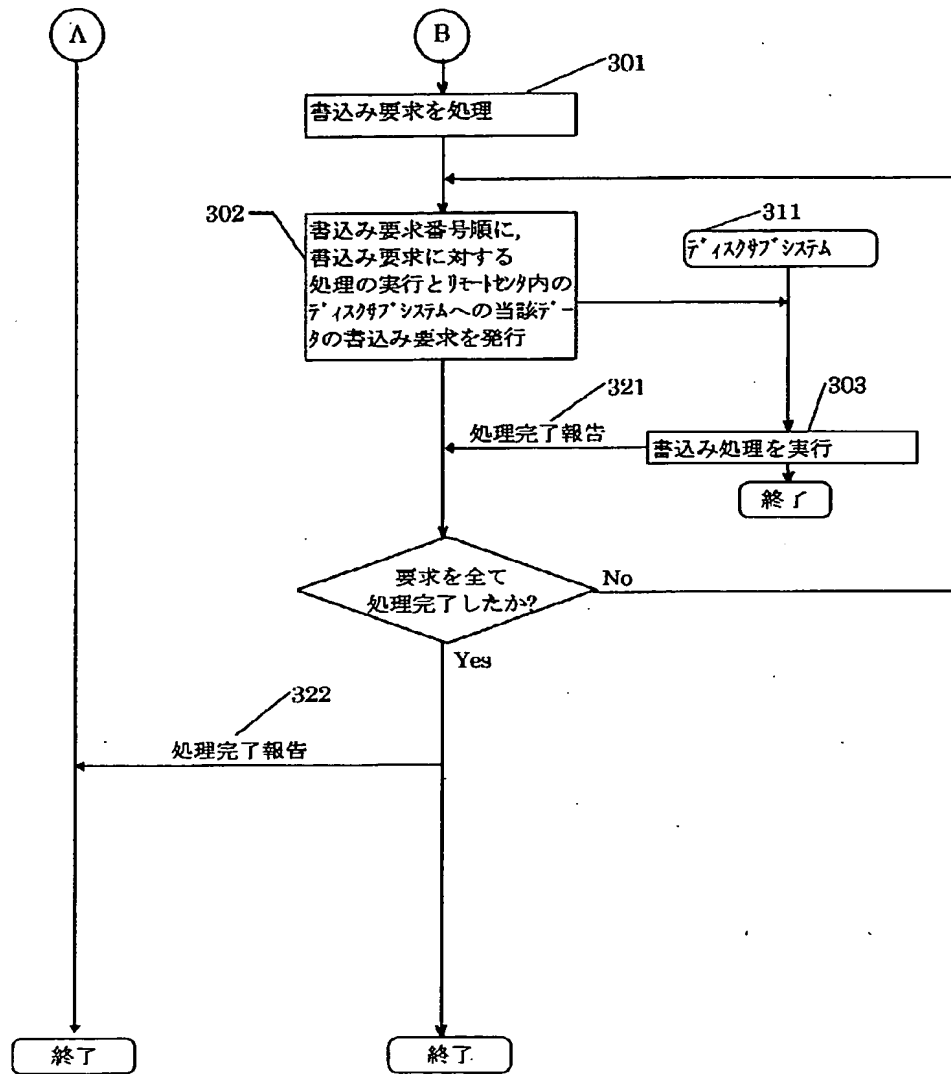
【図2】

図2



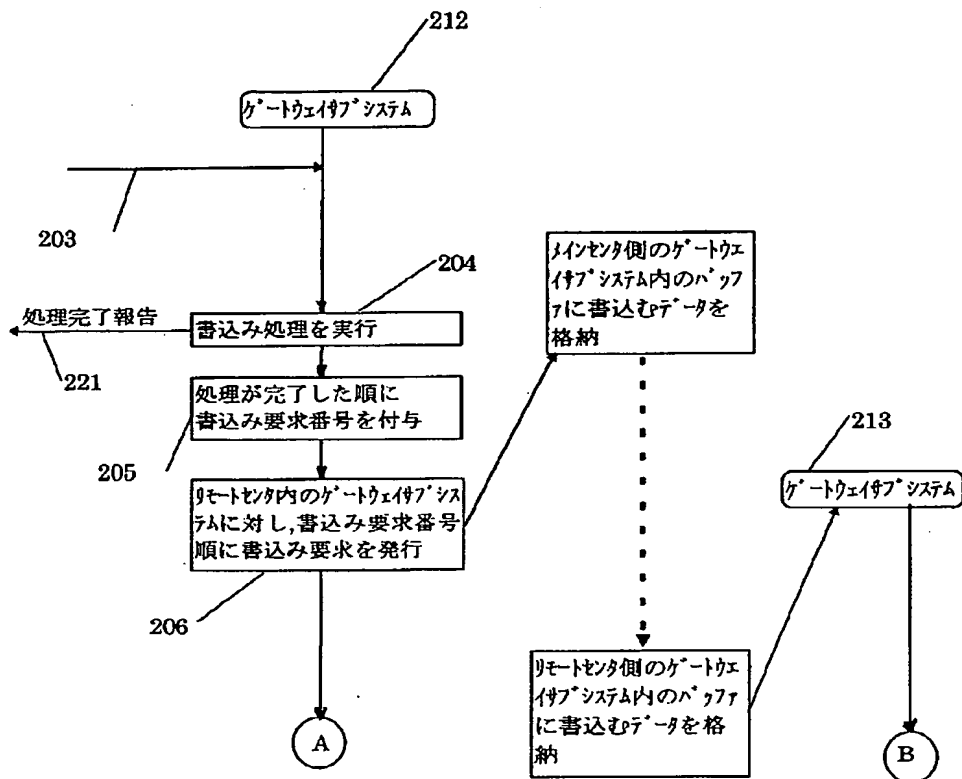
【図3】

図3



【図4】

図4



フロントページの続き

(72)発明者 島田 朗伸
神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内

Fターム(参考) 5B018 GA04 HA05 MA12 QA01
5B034 AA01 BB17 CC02 DD06
5B065 BA01 CA50 CC08 CE22 EA23
EA35
5B082 DA02 DE03 GB02 GB06 HA03
HA10
5B083 AA02 AA09 CD11 CE01 DD13
EE08 GG05

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-244597

(43)Date of publication of application : 19.09.1995

(51)Int.Cl.

G06F 11/20

G06F 3/06

G06F 13/00

(21)Application number : 06-301146

(71)Applicant : INTERNATL BUSINESS MACH.
CORP <IBM>

(22)Date of filing : 05.12.1994

(72)Inventor : KERN ROBERT F
KERN RONALD MAYNARD
MCBRIDE GREGORY EDWARD
SHACKELFORD DAVID MICHAEL

(30)Priority

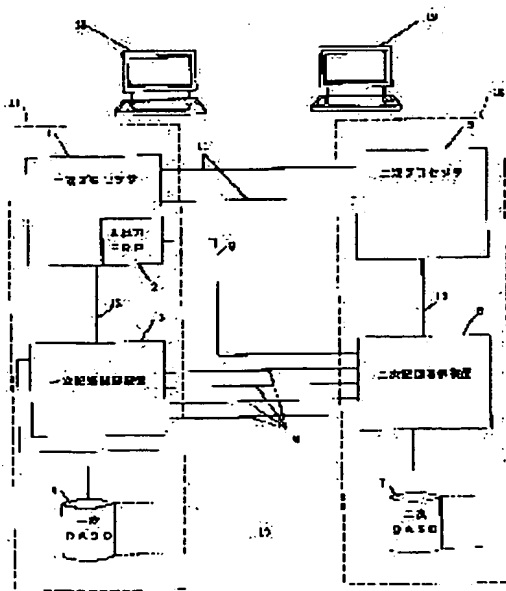
Priority number : 94 199444 Priority date : 22.02.1994 Priority country : US

(54) METHOD FOR FORMING CONSISTENCY GROUP TO PROVIDE DISASTER RECOVERY FUNCTION, AND ITS RELATED SYSTEM

(57)Abstract:

PURPOSE: To provide a remote data shadowing system which provides a real-time disaster recovery function on a storage area base.

CONSTITUTION: A write input-output operation is performed in a storage sub-system on the primary side 14 by record update on the primary side 14. A time-stamp is attached to this write input-output operation and the time, order and physical position of the record update are collected in a primary data mover. The primary data mover divides plural sets of record update and their related control information into groups based on prescribed time intervals, adds a prefix header to the record update and thereby forms a self-description



record set. The self-description record set is sent to a remote secondary side 15, and such a consistency group is formed that the record update is ordered to be able to shadow the record update in the sequence that matches the sequence where the write input-output operation was performed on the primary side 14 by the record update.

LEGAL STATUS

[Date of request for examination] 05.12.1994

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3149325

[Date of registration] 19.01.2001

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 特 許 公 報 (B 2)

(11) 特許番号

特許第3149325号

(P3149325)

(45) 発行日 平成13年3月26日 (2001.3.26)

(24) 登録日 平成13年1月19日 (2001.1.19)

(51) Int.Cl.⁷
G 0 6 F 11/20
3/06
12/00
12/16

識別記号
3 1 0
3 0 4
5 3 3
3 1 0

F I
G 0 6 F 11/20
3/06
12/00
12/16

3 1 0 C
3 0 4 F
5 3 3 J
3 1 0 M

請求項の数10(全 29 頁)

(21) 出願番号 特願平6-301146
(22) 出願日 平成6年12月5日 (1994.12.5)
(65) 公開番号 特開平7-244597
(43) 公開日 平成7年9月19日 (1995.9.19)
審査請求日 平成6年12月5日 (1994.12.5)
(31) 優先権主張番号 1 9 9 4 4 4
(32) 優先日 平成6年2月22日 (1994.2.22)
(33) 優先権主張国 米国 (U S)

(73) 特許権者 390009531
インターナショナル・ビジネス・マシー
ンズ・コーポレーション
INTERNATIONAL BUSI
NESS MACHINES COR
PORATION
アメリカ合衆国10504、ニューヨーク州
アーモンク (番地なし)
(72) 発明者 ロバート・フレデリック・カーン
アメリカ合衆国85730 アリゾナ州ツー
ソン イースト・カレヒコ・ストリート
8338
(74) 復代理人 100065455
弁理士 山本 仁朗 (外2名)

審査官 多賀 実

最終頁に続く

(54) 【発明の名称】 災害復旧機能を提供するために整合性グループを形成する方法および関連するシステム

(57) 【特許請求の範囲】

【請求項1】 一次データ・ムーバおよびレコード更新を発生する複数のアプリケーションを実行する一次プロセッサを備える一次側と、一次プロセッサと通信可能に接続された二次プロセッサを備える二次側とを含む災害復旧のための遠隔データ・シャドーイング・システムであって、一次プロセッサが一次記憶サブシステムに連結され、一次プロセッサから一次記憶サブシステムに発する書込み入出力操作に従ってレコード更新を記憶するための記憶装置が一次記憶サブシステムに備えられ、一次側が一次側における時間依存操作を同期させるための共通システム・タイマを更に備えており、二次側が順序整合性のある順序でデータ更新のコピーを記憶するための二次記憶サブシステムを備えているシステムにおける順序整合性のある順序でデータ更新をシャドーイングする方

法において、前記方法が、

- (a) 一次記憶サブシステムで発生する各書込み入出力操作にタイム・スタンプを付けるステップと、
- (b) 一次記憶サブシステムからレコード更新についてのレコード・セット情報を収集するステップと、
- (c) データ更新およびレコード・セット情報を一次データ・ムーバに書込んでレコード・セットを生成ステップと、
- (d) 二次システムにおける書込み入出力操作の順序を再生成するために二次プロセッサにより使用される自己記述レコード・セットを生成するためレコード・セットの各々にヘッダーを付するステップと、
- (e) 所定の時間間隔に従う時間間隔グループで自己記述レコード・セットを二次プロセッサに転送するステップと、

(f) 自己記述レコード・セットの時間間隔グループから整合性グループを形成するステップであって、一次記憶サブシステムに発せられた書き込み入出力操作の時間順序に基づいて整合性グループ内においてレコード更新が順序づけされる整合性グループを生成するステップと、

(g) 各整合性グループのレコード更新を整合性ある順序で二次記憶サブシステムにシャドーイングするステップと、

(h) ステップ(e)において受取った各自己記述レコード・セットが完全なものであるかどうかを二次側で判定するステップと、
を含む方法。

【請求項2】レコード・セットが二次プロセッサに非同期的に転送されることを特徴とする、請求項1に記載の方法。

【請求項3】ステップ(f)が二次側で行われることを特徴とする、請求項1に記載の方法。

【請求項4】自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落データ更新を再送信するよう二次側が一次側に要求することが、ステップ

(e)にさらに含まれることを特徴とする、請求項1に記載の方法。

【請求項5】各時間間隔グループが完全なものであるかどうかを二次側で判定するステップ(i)をさらに含むことを特徴とする、請求項1に記載の方法。

【請求項6】間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう二次側が一次側に要求することが、ステップ(i)にさらに含まれることを特徴とする、請求項5に記載の方法。

【請求項7】ステップ(b)が、レコード・セット情報において、各レコード更新が格納されている一次記憶装置上の物理的位置を識別することを特徴とする、請求項1に記載の方法。

【請求項8】ステップ(b)が、レコード・セット情報において、セッション内で一次記憶装置に格納された各レコード更新の順序と更新時間を識別することを特徴とする、請求項7に記載の方法。

【請求項9】ステップ(d)が、接頭部ヘッダにおいて、セッションに関する間隔グループ番号と、そこで参照される各レコード更新のグループ内順序を識別することを特徴とする、請求項1に記載の方法。

【請求項10】レコード更新を生成するアプリケーションを実行する一次側を含み、一次側から離れた位置に二次側を有し、二次側がレコード更新をシャドーイングして、一次側に災害復旧を提供する、リアルタイム・データ・シャドーイングを行う非同期遠隔データ二重化システムにおいて、非同期遠隔データ二重化システムが、一次側の時間依存プロセスを同期させるためのシスプレックス・タイマーと、

アプリケーションを実行し、対応するレコード更新用の書き込み入出力操作を出し、一次データ・ムーバをそこに有する、一次側の一次プロセッサと、

各レコード更新ごとに書き込み入出力操作を1つずつ受け取る複数の一次記憶制御装置であって、それぞれの一次記憶制御装置書き込み入出力操作が一次プロセッサによってシスプレックス・タイマーと同期される、複数の一次記憶制御装置と、

対応する書き込み入出力操作に応じて、レコード更新をそこに格納するための複数の一次記憶装置とを含み、

一次データ・ムーバが、各レコード更新ごとに複数の一次記憶制御装置からレコード・セット情報を収集して、所定のグループのレコード・セット情報に接頭部ヘッダを付加し、接頭部ヘッダと所定のレコード・セット情報

グループが自己記述レコード・セットを形成し、各レコード・セット情報が、一次装置アドレス、シリンダ番号およびヘッド番号(CCHH)、レコード更新順序番号、書き込み入出力タイプ、検索指数、セクタ番号、およびレコード更新時間を含み、接頭部ヘッダが、総データ長、操作タイム・スタンプ、時間間隔グループ番号、およびレコード読取り時間を含み、

一次側から自己記述レコード・セットを受け取る二次データ・ムーバを有する二次側の二次プロセッサと、二次プロセッサに連結された複数の二次記憶制御装置

と、
レコード更新を格納する複数の二次記憶装置とをさらに含み、

二次データ・ムーバが、送信された自己記述レコード・セットが完全なものであるかどうかを判定し、自己記述レコード・セットから整合性グループを形成し、複数の一次記憶装置にレコード更新が書き込まれたときの順序に整合する順序で複数の二次記憶装置に書き込むために各整合性グループから得たレコード更新を複数の二次記憶制御装置に出力する、非同期遠隔データ二重化システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、一般的には、災害復旧技法に関し、より具体的には、直接アクセス記憶装置(DASD)データのリアルタイム遠隔コピーのためのシステムに関する。

【0002】

【従来の技術】データ処理に関連して、効率よくアクセスし、修正および再格納できる大量のデータ(またはレコード)を格納するためには、一般に、データ処理システムが必要である。データ記憶域は、効率よくしかも費用効果の高いデータ記憶を行うために、通常、複数のレベルに、すなわち、階層状態に分かれている。第一のレベル、すなわち、最高レベルのデータ記憶域は、通常はダイナミックまたはスタティック・ランダム・アクセス

・メモリ（DRAMまたはSRAM）である電子メモリを含む。電子メモリは、ナノ秒単位でこのようなバイト数のデータにアクセスすることによってそれぞれの回路に数百万バイトのデータを格納できる、半導体集積回路の形態をとる。アクセスが完全に電子的に行われるため、このような電子メモリは最高速のデータ・アクセスを提供する。

【0003】第二レベルのデータ記憶域は、通常、直接アクセス記憶装置（DASD）を含む。DASD記憶域は、磁気ディスクまたは光ディスクなどを含む可能性があり、これらのディスクは、データのビットを構成する「1」と「0」を表す磁氣的または光学的に変質させたマイクロメートル規模のスポットとしてビット単位のデータをディスク表面に格納する。磁気DASDは、残留磁気材料で被覆した1枚または複数枚のディスクを含む。これらのディスクは、保護環境内に回転式に取り付けられている。各ディスクは、多くの同心トラックまたは間隔が詰まった円に分割されている。データは、各トラックに沿って1ビットずつ順次格納される。ヘッド・ディスク・アセンブリ（HDA）として知られているアクセス機構は、一般に、1つまたは複数の読取り書き込みヘッドを含み、ディスクが回転して読取り書き込みヘッドを通過する際にディスクの表面にデータを転送したりディスクの表面からデータを転送するためにトラック間を移動できるよう、各DASDに設けられている。通常、ミリ秒単位（電子メモリより低速の単位）でこのようなデータにアクセスすることによって、DASDはギガバイト規模のデータを格納できる。所望のデータ記憶位置にディスクおよびHDAを物理的に配置する必要があるため、DASDに格納されたデータへのアクセス速度は低下する。

【0004】第三またはそれ以下のレベルのデータ記憶域は、テープまたはテープとDASDライブラリを含む。このレベルの記憶域では、必要なデータ記憶媒体を選択して装填するのにロボットが必要になるため、ライブラリ内のデータへのアクセス速度はさらに低下する。利点は、テラバイト規模のデータ記憶など、データ記憶容量が非常に大きい割に費用が低減される点である。テープ記憶装置は、バックアップ目的で使用されることが多い。すなわち、階層の第二レベルに格納されたデータは、磁気テープで安全に保管するために複製される。テープまたはライブラリに格納したデータへのアクセス速度は、現在、秒単位である。

【0005】データの紛失は、企業にとって破滅的なものになる恐れがあるので、多くの企業ではバックアップのデータ・コピーを用意することが必須になっている。一次記憶レベルで紛失したデータの復旧に要する時間も、復旧に関して考慮すべき重要な項目である。テープまたはライブラリでのバックアップ速度の改善策としては、二重コピーが挙げられる。二重コピーの例として

は、追加のDASDにデータが書き込まれるように追加のDASDを用意する方法がある（ミラーリングと呼ばれる場合もある）。この場合、一次DASDで障害が発生しても、データを得るために二次DASDを頼りにすることができる。この方法の欠点は、必要なDASDの数が二倍になることである。

【0006】記憶装置を二重に設ける必要性を克服するもう1つのデータ・バックアップ方法としては、低価格装置の冗長アレイ（RAID）構成にデータを書き込む方法がある。この方法では、多くのDASDにデータを割り当てるようにデータが書き込まれる。1つのDASDで障害が発生しても、残りのデータとエラー修正手順を使用すれば、紛失したデータを復旧できる。現在では、数種類のRAID構成が使用できる。

【0007】一般に、記憶装置または記憶媒体で障害が発生した場合のデータの復旧には、上記のバックアップ方法で十分である。このようなバックアップ方法では、二次データは一次データのミラーになる、つまり、二次データは一次データと同じボリューム通し番号（VOLSER）とDASDアドレスを持つので、これらの方法は装置障害の場合だけ有用である。しかし、ミラーリングされた二次データを使用しても、システム障害の復旧は行えない。このため、地震、火災、爆発、台風など、システム全体またはシステム使用現場を破壊する災害が発生した場合にデータを復旧するには、さらに保護が必要である。つまり、災害復旧のためには、一次データから離れた位置にデータの二次コピーを格納しておく必要がある。災害保護を行う既知の方法としては、毎日または毎週、テープにデータをバックアップする方法がある。この場合、車両でテープを収集し、通常、一次データ位置から数キロメートル離れた安全な保管区域にテープを移送する。このバックアップ計画には、バックアップ・データを検索するのに何日もかかると同時に、数時間分または数日分のデータが失われ、最悪の場合、同じ災害によって保管場所も破壊されてしまう恐れがあるという問題がある。多少改善されたバックアップ方法としては、毎晩、バックアップ場所にデータを転送する方法が考えられる。この方法では、より離れた遠隔地にデータを格納することができる。しかし、この場合も、二重コピー方法と同様、バックアップが連続して行われるわけではないので、次のバックアップまでにデータの一部が失われる可能性がある。このため、一部のユーザにとっては受け入れられないほど、相当なデータ量が失われる恐れがある。

【0008】さらに最近導入されたデータ災害復旧方法としては、遠隔方式だけでなく、連続方式でもデータがバックアップされる、遠隔二重コピーがある。あるホスト・プロセッサから別のホスト・プロセッサへ、またはある記憶制御装置から別の記憶制御装置へ、あるいはこれらを組み合わせた方法で二重データを送信するには、

このプロセスを実現するために相当な量の制御データが必要になる。しかし、オーバヘッドが高いために、二次側が一次側の処理に遅れないようにする能力が妨げられる可能性があり、このため、災害が発生した場合に二次側が一次側を復旧する能力が脅かされる。

【0009】したがって、最小限の制御データを使用して、一次処理側のデータと一致するデータのリアルタイム更新を行う方法および装置を提供することが望まれている。この方法および装置は、復旧される特定のアプリケーション・データから独立して、つまり、特定のアプリケーション・データをベースとするのではなく、汎用記憶媒体をベースとして機能する。

【0010】

【発明が解決しようとする課題】本発明の目的は、災害復旧のためにDASDデータを二次側にシャドーイング(shadowing)するための改良された設計および方法を提供することにある。

【0011】

【課題を解決するための手段】本発明の第一の実施例によれば、整合性グループを形成するための方法により、遠隔地からの災害復旧機能が提供される。一次プロセッサで実行される1つまたは複数のアプリケーションによって生成されるデータ更新は、一次記憶サブシステムによって受け取られ、この一次記憶サブシステムは、入出力書込み操作により各データ更新がそこに書き込まれるようにする。一次記憶サブシステムは共通タイマによって同期が取られ、一次プロセッサから離れた位置にある二次システムは、災害復旧のために二次側が使用できるように順序整合性のある順序でデータ更新をシャドーイングする。本方法は、(a)一次記憶サブシステムで行われる各書込み入出力操作にタイム・スタンプを付けるステップと、(b)各データ更新ごとに一次記憶サブシステムから書込み入出力操作のレコード・セット情報を収集するステップと、(c)自己記述レコード・セットが一連の書込み入出力操作を再生成するのに十分なものになるように、データ更新とそれぞれのレコード・セット情報から自己記述レコード・セットを生成するステップと、(d)所定の間隔しきい値に基づいて、自己記述レコード・セットを間隔グループにグループ分けするステップと、(e)最も早い操作タイム・スタンプを持つ自己記述レコード・セットの間隔グループとして、第一の整合性グループを選択するステップとを含み、個々のデータ更新が、一次記憶サブシステム内の入出力書込み操作の時間順に基づいて、第一の整合性グループ内で順序付けされる。

【0012】本発明の他の実施例では、一次システムは1つまたは複数のアプリケーションを実行する一次プロセッサを有し、このアプリケーションがレコード更新を生成し、一次プロセッサがそれに基づく自己記述レコード・セットを生成する。それぞれの自己記述レコード・

セットは、一次システムから離れた位置にある二次システムに送られ、二次システムは、リアルタイム災害復旧のために、自己記述レコード・セットに基づいて、順序整合性のある順序でレコード更新をシャドーイングする。一次プロセッサは一次記憶サブシステムに連結され、一次記憶サブシステムはレコード更新を受け取って、入出力書込み操作によって各レコード更新がそこに格納されるようにする。一次プロセッサは、同期を取るためにアプリケーションと一次記憶サブシステムに共通の時間源を提供するためのシスプレックス・クロックを含み、シスプレックス・クロックによって同期が取られる一次データ・ムーバは、各レコード更新ごとに一次データ・ムーバにレコード・セット情報を提供するよう、一次記憶サブシステムに指示する。一次データ・ムーバは、複数のレコード更新とそれに対応する各レコード・セット情報を時間間隔グループにグループ分けし、それに接頭部ヘッダを挿入する。それぞれの時間間隔グループは自己記述レコード・セットを形成する。

【0013】本発明の上記およびその他の目的、特徴、および利点は、添付図面に図示する本発明の実施例に関する以下の詳細な説明から明らかになるだろう。

【0014】

【実施例】一般的なデータ処理システムは、データを計算して操作し、データ機能記憶管理サブシステム/多重仮想記憶システム(DFSMS/MVS)ソフトウェアなどを実行するために、IBMシステム/360またはIBMシステム/370プロセッサなどのホスト・プロセッサの形態をとり、少なくとも1台のIBM 3990記憶制御装置がそれに接続され、その記憶制御装置が、メモリ制御装置と、それに組み込まれた1つまたは複数のタイプのキャッシュ・メモリとを含む場合がある。さらに記憶制御装置は、IBM 3380または3390 DASDなどの1群の直接アクセス記憶装置(DASD)に接続されている。ホスト・プロセッサが実質的な計算能力を提供するのに対し、記憶制御装置は、大規模データベースを効率よく転送し、ステージ(stage)/デステージ(destage)し、変換し、一般的にアクセスするのに必要な諸機能を提供する。

【0015】一般的なデータ処理システムの災害復旧保護では、一次DASDに格納した一次データを二次側または遠隔地でバックアップする必要がある。一次側と二次側との距離は、ユーザが受け入れられる危険のレベルによって決まり、数キロメートルから数千キロメートルの範囲が可能である。二次側または遠隔地は、バックアップ・データ・コピーを提供するだけでなく、一次システムが使用不能になった場合に一次システムの処理を引き継ぐのに十分なシステム情報を持っていなければならない。その理由は、主に、単一の記憶制御装置では一次側と二次側に設けた一次および二次DASDストリングの両方にデータを書き込めないためである。むしろ、一

次記憶制御装置に接続された一次DASDストリングには一次データが格納されるのに対し、二次記憶制御装置に接続された二次DASDストリングには二次データが格納されるのである。

【0016】二次側は、一次側から十分離れている必要があるだけでなく、一次データをリアルタイムでバックアップできなくてはならない。二次側は、一次データが更新されたときに、最小限の遅延で一次データをバックアップする必要がある。しかも、二次側は、一次側で実行され、データまたは更新を生成するアプリケーション・プログラム（たとえば、IMSやDB2）を考慮せずに、一次データをバックアップしなければならない。二次側に要求される難しい課題は、二次データの順序が整合していなければならないことである。つまり、二次データは一次データと同じ順序でコピーされなければならない（順序整合性）、これはシステムについてかなり考慮を要する問題である。順序整合性は、それぞれが1つのデータ処理システム内の複数のDASDを制御する記憶制御装置が複数存在するためにさらに複雑になっている。順序整合性がないと、一次データと一致しない二次データが生成され、その結果、災害復旧が崩壊する恐れがある。

【0017】遠隔データの二重化は、同期と非同期の2つの一般的なカテゴリに分けられる。同期遠隔コピーでは、一次データを二次側に送り、一次DASDの入出力操作を終了する（一次ホストにチャネル終了（CE）と装置終了（DE）を出力する）前にこのようなデータの受取りを確認する必要がある。このため、同期コピーでは、二次側の確認を待っている間に一次DASDの入出力応答時間が遅くなる。一次側の入出力応答遅延は、一次システムと二次システムとの距離に比例して長くなる（これは遠隔距離を数十キロメートル規模に制限する要素である）。しかし、同期コピーは、システム・オーバーヘッドを比較的小さくして、順序が整合したデータを二次側に提供する。

【0018】非同期遠隔コピーでは、二次側でデータが確認される前に一次DASDの入出力操作が完了する（一次ホストにチャネル終了（CE）と装置終了（DE）を出力する）ため、一次側のアプリケーション・システムのパフォーマンスが向上する。このため、一次DASDの入出力応答時間は二次側までの距離に依存せず、一次側から数千キロメートル離れた遠隔地に二次側を設けることもできる。しかし、二次側で受け取ったデータの順序が一次側の更新順序と一致なくなる場合が多いので、データの順序整合性を確保するのに必要なシステム・オーバーヘッドが増加する。したがって、一次側で障害が発生すると、一次側と二次側との間で転送中のデータが一部紛失する恐れがある。

【0019】同期データ・シャドーイング

災害復旧のための同期リアルタイム遠隔コピーでは、コ

ピーされた複数のDASDボリュームが1つのセットを形成する必要がある。さらに、このようなセットを形成するには、各セットを構成するこれらのボリューム（VOLSER）とそれに対応する一次側の同等物を識別するために十分な量のシステム情報を二次側に提供する必要がある。重要なのは、二次側が一次側と「二重対（duplex pair）」を形成するので、1つまたは複数のボリュームがこのセットと同期していない、つまり、「二重対の障害」が発生したときに二次側がそれを認識しなければならない点である。代替経路が再試行される間に一次DASDの入出力が遅延するため、非同期遠隔コピーより同期遠隔コピーの方が、接続障害を認識しやすい。一次側は、二次側用の更新を待ち行列に入れる間、一次側が続行できるようにコピーを中止または中断することができ、一次側は、二次側が同期していないことを示すためにこのような更新にマークを付ける。二次側がいつでも災害復旧に使用できるようにするため、二次側が一次側と同期しなくなる原因になりそうな例外条件を認識することが必要である。エラー条件や復旧処置によって、二次側が一次側と一致しなくなってしまう。

【0020】しかし、二次DASDが存在しアクセス可能である状態で、二次側と一次側の接続を維持しても、内容の同期は確保されない。いくつかの理由から、二次側は一次側との同期性を失う場合もある。二重対が形成されたときに二次側は当初同期しておらず、初期データ・コピーが完了したときに同期に達する。一次側が二次側に更新済みデータを書き込めない場合、一次側は二重対を解除することもある。この場合、一次側は、更新アプリケーションが続行できるように、二重対が中断された状況で一次DASDに更新内容を書き込む。このため、二重対が復元されるまで、一次側は露出状態で、つまり、現行の災害保護コピーを使わずに実行を続ける。二重対が復元されても、二次側は直ちに同期状態になるわけではない。この時点で保留状態の更新を適用した後で、二次側は同期状態に戻る。一次側は、該当ボリュームに関する中止コマンドを一次DASDに出すことによって、二次側の同期を喪失させることもできる。この中止コマンドが終了し、二重対が再確立され、保留状態の更新がコピーされると、二次側は一次側と再同期する。また、オンライン保守でも、同期を喪失させることができる。

【0021】二次ボリュームが一次ボリュームと同期していない場合、二次ボリュームは、二次システムの復旧と一次側アプリケーションの再開に使用することができない。二次側の非同期ボリュームは非同期ボリュームとして識別されなければならない、二次側の復旧引継ぎ手順は、アプリケーション・アクセスを否定する（そのボリュームを強制的にオフラインにするか、そのVOLSERを変更する）ために非同期ボリュームを識別する必要がある。一次側ホストがアクセス不能になった場合に一

次側を復旧するために、二次側が呼び出されることがある。このため、二次側では、すべてのボリュームの同期状態に関するすべての関連情報が必要である。二次記憶サブシステム、すなわち、二次記憶制御装置とDASDは、一次側で検出された例外によって一次側が同期を解除する原因となるすべての条件を判別できるわけではない。たとえば、二次側が把握していない一次入出力経路またはリンクの障害のために一次側が二次側の対等機能にアクセスできない場合、一次側は二重対を解除することがある。この場合、二次側が同期状態を示すのに対し、一次側は、二重対が解除されたことを示す。

【0022】非同期二重対ボリュームが存在することを、外部通信によって二次側に通知することができる。これは、ユーザ・システム管理機能を使用することで認識できる。一次側の入出力操作はチャネル終了/装置終了/装置チェック(CE/DE/UC)状況で終了し、センス・データがエラーの特徴を示す。このような形式の入出力構成の場合、エラー回復プログラム(ERP)がエラーを処理し、入出力の完了を一次側アプリケーションに通知する前に二次プロセッサに適切なメッセージを送る。この場合、ERPの二重対中止メッセージを認識し、その情報を二次側で確保するのは、ユーザの責任である。一次側の代わりに動作可能になるよう二次側が頼りにされている場合は、始動手順によって二次DASDが二次ホストにオンライン接続され、アプリケーションの割振りのために非同期ボリュームがオンライン接続されていないことを確認するために、二次DASDサブシステムに格納された同期状況が検索される。この同期状況をすべてのERP二重対中止メッセージと組み合わせると、二次側の非同期ボリューム全体を示すピクチャが得られる。

【0023】ここで図1を参照して説明すると、同図には、一次側14と二次側15を有し、二次側15が一次側14から20キロメートル離れている、災害復旧システム10が示されている。一次側14は、そこで実行されているアプリケーションと、システム入出力およびエラー回復プログラム2(以下、入出力ERP2という)とを有するホスト・プロセッサまたは一次プロセッサ1を含む。一次プロセッサ1は、DFSMS/MVSオペレーティング・ソフトウェアを実行するIBMエンタープライズ・システム/9000(ES/9000)などでもよく、さらにそこで複数のアプリケーション・プログラムを実行することもできる。一次プロセッサ1には、IBM3990-6型記憶制御装置などの一次記憶制御装置3がチャネル12を介して接続されている。当技術分野で既知の通り、複数のこのような一次記憶制御装置3を1台の一次プロセッサ1に接続するか、あるいは複数の一次プロセッサ1を複数の一次記憶制御装置3に接続することができる。一次記憶制御装置3には、IBM3390DASDなどの一次DASD4が接

続されている。複数の一次DASD4を一次記憶制御装置3に接続することができる。一次記憶制御装置3と、これに接続された一次DASD4によって、一次記憶サブシステムが形成される。また、一次記憶制御装置3と一次DASD4は、単一の一体型ユニットであってもよい。

【0024】二次側15は、チャネル13を介してIBM3990-3型などの二次記憶制御装置6に接続されたIBMES/9000などの二次プロセッサ5を含む。二次記憶制御装置6には、さらにDASD7が接続されている。一次プロセッサ1は、チャネル・リンクまたはT1/T3電話回線リンクなどの少なくとも1つのホスト間通信リンク11によって二次プロセッサ5に接続されている。一次プロセッサ1は、複数のエンタープライズ・システム接続(ESCON)リンク9などによって二次記憶制御装置6との直接接続を確保することもできる。その結果、入出力ERP2は、必要であれば、二次記憶制御装置6と通信可能になる。一次記憶制御装置3は、複数のESCONリンクなどの複数の対等通信リンク8を介して二次記憶制御装置6と通信する。

【0025】一次プロセッサ1で実行されるアプリケーション・プログラムによって書込み入出力操作が実行されると、入出力操作が正常に完了したことを示すハードウェア状況としてチャネル終了/装置終了(CE/DE)が出力される。入出力操作が正常に完了すると、一次プロセッサ1のオペレーティング・システム・ソフトウェアは、そのアプリケーションに書込み入出力成功のマークを付け、それにより、アプリケーション・プログラムは、最初または前の書込み入出力操作の正常終了に依存する可能性のある次の書込み入出力操作に移行できるようになる。これに対して、書込み入出力操作が不成功に終わった場合は、チャネル終了/装置終了/装置チェック(以下、CE/DE/UCという)という入出力状況が一次プロセッサ1のオペレーティング・システム・ソフトウェアに出力される。装置チェックを出力した後、入出力ERP2は、制御権を引き継ぎ、失敗した書込み入出力操作の特徴に関する具体的なセンス情報を一次記憶制御装置3から入手する。あるボリュームに固有のエラーが発生した場合は、そのエラーに関連する固有の状況が入出力ERP2に出力される。その後、入出力ERP2は、一次記憶制御装置3と二次記憶制御装置6との間、または最悪の場合は、一次プロセッサ1と二次プロセッサ5との間のデータ保全性を維持するために、新たな対等通信同期エラー回復を実行することができる。

【0026】図2および図3を参照して説明すると、同図にはエラー回復手順が示されている。図2のステップ201は、一次プロセッサ1で実行されるアプリケーション・プログラムが一次記憶制御装置3にデータ更新を送信することを含む。ステップ203では、そのデータ

更新が一次DASD 4に書き込まれ、そのデータ更新が二次記憶制御装置6にシャドーイングされる。ステップ205では、二重対の状況がチェックされ、一次側と二次側が同期しているかどうかが判別される。二重対の状況が同期状態になっている場合、ステップ207でデータ更新が二次DASD 7に書き込まれ、一次プロセッサ1での処理は、そこで実行されるアプリケーション・プログラムを介して続行される。

【0027】二重対が「障害発生」状態になっている場合、ステップ209で一次記憶制御装置3は、二重対で中断または障害が発生していることを一次プロセッサ1に通知する。二重対は、通信リンク8による一次記憶制御装置3と二次記憶制御装置6との通信障害によって「障害発生」状態になる場合がある。あるいは、二重対は、一次サブシステムまたは二次サブシステムいずれかのエラーによって「障害発生」状態になる場合もある。障害が通信リンク8で発生している場合、一次記憶制御装置3は、二次記憶制御装置6に直接、障害を連絡することができない。そこで、一次記憶制御装置3は、ステップ211で入出力状況としてCE/DE/UCを一次プロセッサ1に返す。入出力ERP 2は、アプリケーション・プログラムを静止させ、書き込み入出力操作を要求するアプリケーションに制御権を返す前に、エラー回復とデータ保全性のためにステップ213で一次プロセッサ1の制御権を引き継ぐ。

【0028】図3は、入出力ERP 2が実行する諸ステップを表している。ステップ221で入出力ERP 2は一次記憶制御装置3にセンス入出力を出す。センス入出力操作は、入出力エラーの原因を記述する情報を返す。すなわち、このデータ記述情報は、具体的なエラーに関して記憶制御装置または二重対の操作に固有のものになる。一次記憶制御装置3と二次記憶制御装置6との間の対等通信リンク8で障害が発生したことがデータ記述情報によって示された場合、ステップ223で入出力ERP 2は、一次記憶制御装置3および二次記憶制御装置6に対して、関係ボリュームを同期遠隔コピー障害状態に入れるように指示する記憶制御装置レベル入出力操作を出す。この二次記憶制御装置6は、複数のESCONリンク9またはホスト間通信リンク11を介して入出力ERP 2から関係ボリュームの状態を受け取ることができる。その結果、二重対操作の現在の状況は、一次プロセッサ1で実行されるアプリケーションとともに、一次プロセッサ1および二次プロセッサ5の両方で維持される。コンソール18および19は、それぞれ一次プロセッサ1および二次プロセッサ5からの情報をやりとりするために設けられ、入出力ERPは、両方のコンソール18および19に状況情報を通知する。

【0029】一次記憶制御装置3および二次記憶制御装置6への同期遠隔コピー入出力操作障害が正常に完了したとき、ステップ225ではデータ保全性が維持されて

いる。このため、二次側15で復旧を試みると、二次記憶制御装置6は、「同期遠隔コピー障害」というマークを付けたボリュームを、データ回復手段（ボリューム上のそのデータの状態を判別するための従来のデータベース・ログまたはジャーナル）によってそのボリュームのデータとその同期グループ内の他のデータとの同期が取られるまで使用できないものとして識別する。

【0030】ステップ227では、同期遠隔コピー障害の状況更新について一次記憶制御装置3と二次記憶制御装置6で行われた入出力操作の正常終了を入出力ERP 2が受け取ったかどうかを判別するテストが行われる。正常終了すると、入出力ERP 2は、ステップ229で一次プロセッサ1に制御権を返す。正常終了していない場合は、ステップ231で次のレベルの復旧通知が行われる。この通知には、障害発生ボリュームと、一次記憶制御装置3または二次記憶制御装置6のいずれかのそのボリュームの状況が正しくない可能性があることを、コンソール18を介してオペレータに通知することが含まれる。この通知は、そこで具体的なボリューム状況を示すために、コンソール19または共用DASDデータ・セットを介して二次側15にシャドーイングされる。

【0031】ステップ233で、エラー・ログ記録データ・セットが更新される。この更新は、一次DASD 4または他の記憶場所のいずれかに書き込まれ、二次側15にシャドーイングされる。このエラー回復処置が完了すると、入出力ERP 2はステップ235で、書き込み入出力操作障害に関する「永続エラー」回復を一次側アプリケーションに実行させるために、一次側アプリケーションの書き込み入出力操作に「永続エラー」を通知する。エラーが修正されると、ボリューム状態は、まず保留状態（変更データの再コピー）に回復し、次に全二重に回復することができる。その後、二重対が再確立されると、データを二次DASD 7に再適用することができる。

【0032】二重対を確立する場合、顧客の要求に応じて、ボリュームをCRITICALと識別することができる。CRITICALボリュームの場合、ある操作の結果、二重対の障害が発生すると、実際のエラー箇所とは無関係に、一次ボリュームの永続エラー障害が報告される。CRIT=Yの場合、障害発生対の一次DASD 406に書き込もうとするその後のすべての試みは、永続エラーを受け取ることになり、対をなす二次ボリュームにシャドーイングできないデータは、その一次ボリュームに一切書き込まれなくなる。このため、必要であれば、一次側アプリケーションの処置および入出力データ操作との完全同期が可能になる。

【0033】その結果、本明細書に記載する災害復旧システム10では、入出力命令（チャネル・コマンド・ワード（CCW））を有する一次ホスト処理エラー回復手順によって、一次および二次同期遠隔コピー・ボリュー

ムの状況を二重対から障害発生二重対へ変更できるようにする同期遠隔コピーを取り入れ、それにより、複数タイプの一次および二次サブシステム・エラーの場合にデータ保全性を維持する。アプリケーション・ベースのバックアップではなく、データ更新がリアルタイムで複写される記憶域ベースのバックアップが設けられている。また、災害復旧システム10は、(1)一次および二次記憶制御装置ボリューム状況更新、(2)オペレータ・メッセージまたはエラー・ログ記録共通データ・セットを介して具体的なボリューム更新状況に関して一次および二次ホスト・プロセッサが通知すること、および(3)CRITICALボリューム識別などの、複数レベルの一次/二次状況更新を試み、ボリューム対が障害発生二重対になる場合は、一次ボリュームへのその後の更新を防止することができる。このため、リアルタイムの完全エラー災害復旧が達成される。

【0034】非同期データ・シャドーイング

非同期遠隔データ・シャドーイングは、1回の災害で一次側と二次側の両方が崩壊してしまう確率を低減するために一次側と二次側との距離をさらに大きくする必要がある場合、または一次側アプリケーションのパフォーマンスへの影響を最小限に抑える必要がある場合に使用する。一次側と二次側との距離は、現在では地球全体またはそれ以上に延長できるが、複数の一次サブシステムの背後にある複数のDASDボリュームにわたる書込み更新を複数の二次サブシステムに同期させることは、さらに複雑である。二次記憶サブシステム上でシャドーイングするために、一次データ・ムーバを介して一次記憶制御装置から二次データ・ムーバへレコード書込み更新を発送することができるが、両者間でやりとりされる制御データの量は、最小限でなければならず、同時に、複数の記憶制御装置に隠れている複数のDASDボリュームにわたる一次システムの場合と同様、複数の記憶制御装置にわたる二次システム上でレコード書込み更新の順序を正確に再構築できるものでなければならない。

【0035】図4は、一次側421と遠隔側または二次側431とを含む非同期災害復旧システム400を示している。一次側421は、DFSMS/MVSホスト・ソフトウェアを実行するIBM ES/9000などの一次プロセッサ401を含む。一次プロセッサ401は、IMSおよびDBSアプリケーションなどのアプリケーション・プログラム402および403と、一次データ・ムーバ(PDM)404をさらに含む。一次プロセッサ401には、そこで実行されるすべてのアプリケーション(402、403)に共通の基準を提供するために、共通シプレックス・クロック(sysplex clock)407が設けられ、すべてのシステム・クロックまたは時間源(図示せず)がシプレックス・クロック407に同期し、すべての時間依存プロセスが相互に正しいタイミングで動作するようになっている。たとえば、

一次記憶制御装置405は、単一の一次記憶制御装置406への2回の連続する書込み入出力操作が同じタイム・スタンプ値を示さないように、複数のレコード書込み更新時間を確実に区別するのに適した解像度に同期している。シプレックス・クロック407の解像度(正確さではない)は重要である。PDM404は、シプレックス・クロック407に接続された状態で図示されているが、書込み入出力操作がそこで発生するわけではないので、シプレックス・クロック407に同期させる必要はない。また、一次プロセッサ401が単一の時間基準(たとえば、単一のマルチプロセッサES/9000システム)を有する場合には、シプレックス・クロック407は不要である。

【0036】一次プロセッサ401には、IBM 3990-6型記憶制御装置などの複数の一次記憶制御装置405が光ファイバ・チャネルなどの複数のチャネルを介して接続されている。また、各一次記憶制御装置405には、IBM 3390DASDなどの複数の一次DASD406からなる少なくとも1つのストリングが接続されている。一次記憶制御装置405と一次DASD406によって、一次記憶サブシステムが形成される。各記憶制御装置405と一次DASD406は、個別のユニットである必要はなく、両者を組み合わせて単一のドロウにしてもよい。

【0037】一次側421から数千キロメートル離れた位置に配置される二次側431は、一次側421と同様に、そこで動作する二次データ・ムーバ(SDM)414を有する二次プロセッサ411を含む。あるいは、一次側と二次側が同じ場所に存在してもよく、さらに、一次データ・ムーバと二次データ・ムーバが単一のホスト・プロセッサに常駐してもよい(二次DASDは防火壁のすぐ上に設けてもよい)。二次プロセッサ411には、当技術分野で既知の通り、光ファイバ・チャネルなどのチャネルを介して複数の二次記憶制御装置415が接続されている。記憶制御装置415には、複数の二次DASD416と1つの制御情報DASD417(複数可)が接続されている。記憶制御装置415とDASD416および417によって、二次記憶サブシステムが構成される。

【0038】一次側421は、通信リンク408を介して二次側431と通信する。より具体的には、一次プロセッサ401は、仮想記憶通信アクセス方式(VTAM)通信リンク408などの通信プロトコルによって、二次プロセッサ411にデータと制御情報を転送する。この通信リンク408は、電話(T1、T3回線)、無線、無線/電話、マイクロ波、衛星などの複数の適当な通信方式によって実現できる。

【0039】非同期データ・シャドーイング・システム400は、一次DASD406へのすべてのデータ書込みの順序が保持され、二次DASD416に適用される

(すべての一次記憶サブシステムにわたるデータ書き込み順序を保持する)ように、一次記憶制御装置405から制御データを収集する機能を含む。二次側431に送られるデータおよび制御情報は、データ保全性を保持するのに一次側421の存在が不要になるほど、十分なものでなければならない。

【0040】アプリケーション402、403は、データまたはレコード更新を生成するが、このレコード更新は、一次記憶制御装置405によって収集され、PDM404によって読み取られる。それぞれの一次記憶制御装置405は、非同期遠隔データ・シャドーイング・セッションのためにそれぞれのレコード更新をグループ化し、非特定一次DASD406のREAD要求を介してPDM404にこれらのレコード更新を提供する。一次記憶制御装置405からPDM404へのレコード更新の転送は、START入出力操作の回数および読取りから読取りの遅延を最小限にしながら、各一次記憶制御装置405と一次プロセッサ401との間で転送されるデータの量を最大にするように、PDM404によって制御され、最適化される。PDM404は、非特定READ間の時間間隔を変えることで、一次記憶制御装置とホストとのこの最適化だけでなく、二次DASD416用のレコード更新の通用期間も制御することができる。

【0041】データ保全性を維持しながら、PDM404がレコード更新を収集し、そのレコード更新をSDM414に送信するには、すべての一次記憶サブシステムにおいて二次DASD416に対して行われる一次DASD406のレコードWRITEシーケンスを再構築するのに十分な制御データとともに、特定の期間の間、適切な複数の時間間隔でレコード更新を送信する必要がある。一次DASD406のレコードWRITEシーケンスの再構築は、自己記述レコードをPDM404からSDM414に渡すことによって達成される。SDM414は、所与の時間間隔分のレコードが紛失しているかどうか、または不完全になっているかどうかを判別するために、その自己記述レコードを検査する。

【0042】図5および図6は、接頭部ヘッダ500(図5)と、一次記憶制御装置405によって生成されたレコード・セット情報600(図6)とを含む、各自記述レコードごとにPDM404が作成するジャーナル・レコード形式を示している。各自記述レコードは、それぞれの時間間隔の時間順に二次DASD416に適用できるように、それぞれの時間間隔ごとにさらにSDM414によってジャーナル処理される。

【0043】ここで図5を参照して説明すると、各レコード・セットの先頭に挿入される接頭部ヘッダ500は、接頭部ヘッダ500と、各レコード・セットごとにSDM414に送信される実際の一次レコード・セット情報600との長さの合計を記述するための総データ長501を含む。操作タイム・スタンプ502は、PDM

404が現在処理している操作セットの開始時間を示すタイム・スタンプである。この操作タイム・スタンプ502は、1組の一次記憶制御装置405に対してREAD RECORD SET機能を実行する際に(シスプレックス・クロック407に応じて)PDM404によって生成される。一次DASD406の書き込みの入出力時間610(図6)は、各一次記憶制御装置405のREAD RECORD SETごとに固有のものである。操作タイム・スタンプ502は、すべての記憶制御装置で共通のものである。

【0044】READ RECORD SETコマンドは、PDM404によって出されるが、以下の条件のいずれかの場合に予測できる。

(1) 一次記憶制御装置405の所定のしきい値に基づく、その一次記憶制御装置のアテンション割込み

(2) 所定の時間間隔に基づく、一次プロセッサ401のタイマ割込み

(3) レコード・セット情報が、使用可能であるがまだ読み取られていない未解決のレコード・セットに関する追加情報を示す場合

条件(2)では、タイマ間隔を使用して、低レベル活動の期間中に二次システム431がどの程度遅れて実行するかを制御する。条件(3)は、PDM404が一次記憶制御装置405の活動に遅れないようにするためにさらに活動を駆動する処理間隔中に、PDM404がすべてのレコード・セットを待ち行列処理しなかった場合に発生する。

【0045】時間間隔グループ番号503は、現行レコード・セット(整合性グループのうちの所与の時間間隔グループについてすべての一次記憶制御装置405にわたるレコードのセット)が属す時間間隔(操作タイム・スタンプ502とレコード読取り時間507によって境界が示される)を識別するためにPDM404が出力する。グループ内順序番号504は、所与の時間間隔グループ503内の各レコード・セットごとに一次記憶制御装置405用のアプリケーションWRITE入出力の順序を(PDM404に対して)識別するためにハードウェアが提供するIDに基づいて導出される。一次SSID(補助記憶域ID)505は、各レコード・セットごとに一次記憶制御装置405の特定の一次記憶制御装置を明確に識別するものである。二次ターゲット・ボリューム506は、パフォーマンス上の考慮事項に応じて、PDM404またはSDM414のいずれかによって割り当てられる。レコード読取り時間507は、すべての一次記憶制御装置405に共通の操作タイム・スタンプを提供し、現行間隔のレコード・セットの終了時間を示す。

【0046】操作タイム・スタンプ502およびレコード読取り時間507は、各一次記憶制御装置405から得た複数組の読取りレコード・セットをグループ分けす

るためにPDM404が使用する。複数組の読取りレコード・セットをグループ分けするための同期はPDM404にだけ重要であるため、PDM404は、シスプレックス・クロック407に接続されていないPDM404だけを動作させる中央処理装置(CPU)クロックに同期してもよい。PDM404はレコード更新を書き込まないが、前述の通り、レコード更新は共通の時間源に同期していなければならない。

【0047】次に図6を参照して説明すると、レコード・セット情報600は、一次記憶制御装置405によって生成され、PDM404によって収集される。更新固有情報601~610は、レコード更新が行われた実際の一次DASD406を含む各レコードの一次装置ユニット・アドレス601を含む。シリンドラ番号/ヘッド番号(CCHH)602は、各レコード更新ごとの一次DASD406上の位置を示す。一次記憶制御装置のセッションIDである一次SSID603は、一次SSID505と同じものである。状況フラグ604は、特定のデータ・レコード620が後に続くかどうかに関する状況情報を提供する。順序番号605および630は、レコード・セット全体(PDM404に転送されたすべてのデータ)が読み取られたかどうかを示すために各レコードに番号を1つずつ割り当てる。一次DASD書き込み出力タイプ606は、各レコードについて行われた書き込み操作のタイプを識別する操作標識である。この操作標識は、更新書き込み、フォーマット書き込み、部分トラック・レコード・フォロー、完全トラック・データ・フォロー、消去コマンド実行、または全書き込み実行を含む。検索指数607は、最初に読み取られたレコード・セット・データ・レコード620に関する初期位置決め情報を示す。セクタ番号608は、レコードが更新されたセクタを識別する。カウント・フィールド609は、後続の特定のレコード・データ・フィールド620の数を記述する。一次DASD406の書き込み更新が行われたホスト・アプリケーション時間は、更新時間610に記録される。特定のレコード・データ620は、各レコード更新ごとのカウント/キー/データ(CKD)フィールドを提供する。最後に、順序番号630は、読み取られたレコード・セット全体がPDM404に転送されたかどうかを示すために順序番号605と比較される。

【0048】一次DASD406でレコード更新が書き込まれたのと同じ順序でSDM414がそのレコード更新をコピーできるように、ソフトウェア・グループが呼び出した整合性グループで更新レコードが処理される。整合性グループを作成するのに使用する情報(すべての記憶制御装置405から収集したすべてのレコード・セットにわたる)は、操作タイム・スタンプ502、時間間隔グループ番号503、グループ内順序番号504、一次制御装置SSID505、レコード読取り時間507、一次装置アドレス601、一次SSID603、お

よび状況フラグ604を含む。1つの時間間隔グループ用のすべてのレコードがSDM414側で各記憶制御装置405ごとに受け取られたかどうかを判別するのに使用する情報は、時間間隔グループ番号503、グループ内順序番号504、物理制御装置ID505、および一次SSID603および各操作時間間隔ごとに各一次記憶制御装置405から返される読取りレコード・セットの総数を含む。完全復旧可能な一次DASD406のレコード更新と同等に二次DASD416上にレコード更新を配置するのに必要な情報は、二次ターゲット・ボリューム506、CCHH602、一次DASD書き込み出力タイプ606、検索指数607、セクタ番号608、カウント609、更新時間610、および特定のレコード・データ620を含む。

【0049】図7および図8は、復旧時間とジャーナル転送時間を単純化した現行ジャーナル内容を記述するための状態テーブル700とマスタ・ジャーナル800をそれぞれ示す。状態テーブル700は、PDM404とSDM414が収集し、両者に共通の構成情報を提供し、一次記憶制御装置のセッションID(SSID番号)およびその制御装置でのボリュームと、対応する二次記憶制御装置のセッションIDおよび対応するボリュームとを含む。このため、構成情報は、どの一次ボリューム710または一次DASDエクステントが二次ボリューム711または二次DASDエクステントにマッピングされるかを追跡する。状態テーブル700まで単純に拡張して部分ボリューム・エクステント712(CCHHからCCHHまで)を示す場合、部分ボリューム遠隔コピーは、ここに記載するのと同じ非同期遠隔コピー方法を使用して達成できるが、完全ボリュームの場合より細分性(トラックまたはエクステント)はより細くなる。

【0050】マスタ・ジャーナル800は、整合性グループ番号、ジャーナル・ボリューム上の位置、および操作タイム・スタンプを含む。また、マスタ・ジャーナル800は、整合性グループにグループ化した特定のレコード更新を維持する。状態テーブル700とマスタ・ジャーナル800は、災害復旧をサポートするため、一次システム410がもはや存在しないスタンドアロン環境で動作できなければならない。

【0051】制御項目全体が正しく書き込まれるようにするため、タイム・スタンプ制御は各マスタ・ジャーナル800の前後に置かれる。このタイム・スタンプ制御は、さらに二次DASD417に書き込まれる。制御要素は二重項目(1)および(2)を含み、次に示す例のように一方の項目が必ず現行項目になる。

- (1) タイム・スタンプ制御 | 制御情報 | タイム・スタンプ制御
- (2) タイム・スタンプ制御 | 制御情報 | タイム・スタンプ制御

いかなる時点でも、(1)または(2)のいずれかの項目が現行または有効項目になるが、有効項目は前後に等しいタイム・スタンプ制御を持つ項目である。災害復旧では、制御情報を得るために、最新のタイム・スタンプを持つ有効項目を使用する。この制御情報は、状態情報(記憶制御装置、装置、および適用される整合性グループに関する環境情報)とともに、二次記憶制御装置415にどのレコード更新が適用されたかを判別するのに使用する。

【0052】整合性グループ

所定の時間間隔の間にすべての一次記憶制御装置405にわたるすべての読取りレコード・セットが二次側431で受け取られると、SDM414は、受け取った制御情報を解釈し、レコード更新が最初に一次DASD406上で書き込まれたのと同じ順序でそのレコード更新が適用されるように、受け取った読取りレコード・セットをレコード更新・グループとして二次DASD416に適用する。このため、一次側アプリケーションの順序(データ保全性)整合性はすべて二次側431で維持される。このプロセスは、以下、整合性グループの形成と呼ぶ。整合性グループの形成は、次のような仮定に基づいて行われる。(A)独立しているアプリケーション書込みが制御装置の順序命令に違反しない場合は、どのような順序でもアプリケーション書込みを実行できる。

(B)従属しているアプリケーション書込みは、タイム・スタンプの順に実行しなければならないため、アプリケーションは、書込み番号1から制御装置終了、装置終了を受け取る前に従属書込み番号2を実行することができない。(C)第二の書込みは必ず(1)遅いタイム・スタンプを持つ第一の書込みと同じレコード・セット整合性グループに入るか、(2)後続のレコード・セット整合性グループに入る。

【0053】図9を参照して説明すると、同図には、記憶制御装置SSID1、SSID2、およびSSID3など(記憶制御装置はいくつでも含めることができるが、この例では明解にするため3つ使用する)に関する整合性グループの形成例(整合性グループは一次側421または二次側431のいずれにも形成できるはずである)が示されている。時間間隔T1、T2、およびT3は昇順に発生するものと想定する。時間間隔T1の操作タイム・スタンプ502は、記憶制御装置SSID1、SSID2、およびSSID3について設定されている。PDM404は、時間間隔T1〜T3の間に記憶制御装置SSID1、2、および3からレコード・セット・データを入手する。時間間隔T1のSSID1、2、および3に関するレコード・セットは、時間間隔グループ1であるG1(時間間隔グループ番号503)に割り当てられる。グループ内順序番号504は、SSID1、2、および3のそれぞれについて示され、この場合、SSID1は11:59、12:00、および1

2:01に3つの更新を持ち、SSID2は12:00および12:02に2つの更新を持ち、SSID3は11:58、11:59、および12:02に3つの更新を持つ。時間間隔T2およびT3のレコード・セットは列挙されているが、簡略化のため、更新時間の例は示されていない。

【0054】ここで、二次側431で受け取った制御情報およびレコード更新に基づいて、整合性グループNを生成することができる。時間間隔グループ番号1のレコード更新が時間間隔グループ番号2のレコード更新より遅くならないようにするため、記憶制御装置SSID1、2、および3のそれぞれの最後のレコード更新の最も早い読取りレコード・セット時間と等しい、最小時間が設定される。この例では、最小時間は12:01になる。最小時間と等しいかそれ以上の読取りレコード・セット時間を有するレコード更新はすべて整合性グループN+1に含まれる。1つのボリュームに対する2つのレコード更新時間が等しい場合、シスプレックス・クロック407の十分な解像度が与えられる可能性はほとんどないが、時間間隔グループN内の早い順序番号を持つレコード更新は、整合性グループN用のそのグループとともに保管される。ここで、レコード更新は、読取りレコード・セット時間に基づいて順序づけされる。複数のレコード更新の時間が等しい場合、小さい順序番号を持つレコード更新は、大きい順序番号を持つレコード更新の前に置かれる。これに対して、複数のレコード更新のタイム・スタンプが等しいが、ボリュームが異なる場合は、そのレコード更新が同じ整合性グループに保管されている限り、任意の順序にすることができる。

【0055】一次記憶制御装置405が、指定の時間間隔の間に読取りレコード・セットへの応答を完了しなかった場合、その一次記憶制御装置405が完了するまで、整合性グループを形成することはできない。一次記憶制御装置405がその操作を完了しなかった場合は、未着割込みのために、システムの未着割込みハンドラが制御権を受け取り、操作が終了する。これに対して、一次記憶制御装置405が適切な時間に操作を完了した場合は、入出力が完了に至り、通常操作が続行される。整合性グループの形成では、一次記憶制御装置405に対する書込み操作にタイム・スタンプが付けられると予想される。しかし、プログラムによっては、タイム・スタンプが付けられずに書込みが生成されるものもある。この場合、一次記憶制御装置405は、タイム・スタンプとしてゼロを返す。整合性グループの形成は、データが読み取られたタイム・スタンプに基づいて、タイム・スタンプを持たないこれらのレコードの境界を示すことができる。整合性グループの時間別にレコード更新の境界を容易に示せないほど、タイム・スタンプを持たないレコード更新が一定の時間間隔の間に多数発生した場合、二重ボリュームが同期していないというエラーが発生す

る可能性がある。

【0056】図10および図11は、整合性グループを形成する方法を示す流れ図である。図10を参照して説明すると、このプロセスは、ステップ1000から始まり、一次側421が、行うべき遠隔データ・シャドーイングを確立する。ステップ1010では、シスプレックス・クロック407を同期クロック（図4）として使用して、すべてのアプリケーション入出力操作にタイム・スタンプが付けられる。PDM404は、ステップ1020で各一次記憶制御装置405との遠隔データ・シャドーイング・セッションを開始するが、このステップは、データまたはレコードがシャドーイングされる一次ボリュームの識別を含む。ステップ1030では、各アプリケーションWRITE入出力操作（図6を参照）ごとに一次記憶制御装置405によってレコード・セット情報600がトラッピングされる。

【0057】ステップ1040は、前述の通り、アテンション・メッセージを含むプロンプト、所定のタイミング間隔、または読取りレコード数増加の通知に応じて、PDM404が、捕捉したレコード・セット情報600を各一次記憶制御装置405から読み取ることを含む。ステップ1050でPDM404がレコード・セットの読取りを開始すると、PDM404は、各レコード・セットの前に特定のジャーナル・レコード（ジャーナル・レコードは、接頭部ヘッダ500と、レコード・セット情報600を含む）を作成するための接頭部ヘッダ500（図5を参照）を付ける。このジャーナル・レコードには、二次側431（または一次側421）で整合性グループを形成するのに必要な制御情報（およびレコード）が含まれる。

【0058】ステップ1060では、PDM404が通信リンク408を介して（整合性グループがそこで形成される場合は、同じデータ・ムバ・システム内で）SDM414に生成したジャーナル・レコードを送信する。SDM414は、ステップ1070で状態テーブル700を使用し、データ・シャドーイング・セッション用に確立した各時間間隔グループおよび一次記憶制御装置405ごとに、受け取ったレコード更新をグループ番号別および順序番号別に収集する。ステップ1080でSDM414は、ジャーナル・レコードを検査し、各時間間隔グループごとにすべてのレコード情報を受け取ったかどうかを判別する。ジャーナル・レコードが不完全な場合は、ステップ1085によって、SDM414はPDM404に必要なレコード・セットを再送信するよう通知する。PDM404が正しく再送信できない場合は、二重ボリューム対に障害が発生している。ジャーナル・レコードが完全な場合は、SDM414による整合性グループの形成を含むステップ1090が実行される。

【0059】図11を参照すると、同図には、整合性グ

ループを形成するためのステップ1090（図10）を表すステップ1100～1160が示されている。整合性グループの形成は、ステップ1100から始まるが、このステップでは、各ソフトウェア整合性グループが二次DASD417（図4）上のSDM414ジャーナル・ログ（"hardened"）に書き込まれる。ステップ1110は、時間間隔グループが完全かどうかを判別するテストを実行する。すなわち、各一時記憶制御装置405は、少なくとも1つの読取りレコード・セット・バッファを提示したか、レコード・セット・バッファ内にこのようなレコード更新が置かれていないという確認をPDM404から受け取らなければならない。しかもデータ（またはヌル）を持つすべての読取りレコード・セット・バッファがSDM414によって受け取られていなければならない。時間間隔グループが不完全な場合は、ステップ1110は、必要なデータが受け取られるまで、一次記憶制御装置405からのレコード・セットの読取りを再試行する。エラーが発生した場合は、特定の1つまたは複数の二重ボリューム対に障害が発生している可能性がある。完全な時間間隔グループを受け取ると、ステップ1120は、第一の整合性グループ・ジャーナル・レコードを判別する。この第一（または現行）の整合性グループ・ジャーナル・レコードとは、最も早い操作タイム・スタンプ502と、同じ操作タイム・スタンプ502を持つすべてのレコードの最も早い更新時間610を含むレコードである。

【0060】ステップ1130は、現行整合性グループ・ジャーナル・レコードに含まれるレコードを検査して、どのレコードがそこに最後に含めるレコードかを判別する（一部のレコードは除去され、次の整合性グループ・ジャーナル・レコードに含まれる）。現行整合性グループ・ジャーナル・レコードの最後のレコードは、各一次記憶制御装置405ごとに最大更新時間のうちの最小更新時間（最小時間）として判別される（つまり、各一次記憶制御装置405の最後の更新が比較され、これらのうちの最も早いものだけが現行整合性グループ・ジャーナル・レコードに残る）。

【0061】現行整合性グループ・ジャーナル・レコードに残っているこれらのレコード更新は、ステップ1140で、更新時間610とグループ内順序番号504に応じて順序付けされる。レコード更新を持たない一次記憶制御装置405は、整合性グループに関与しない。ステップ1150では、現行整合性グループに残っているレコード更新（最小時間より遅い更新時間を持つもの）が、次の整合性グループに渡される。それぞれのグループ内順序番号504は、空バッファで終わり、その操作時間間隔の間にすべての読取りレコード・セットが読み取られたことを示すはずである。空バッファがない場合は、現行ソフトウェア整合性グループ内の最後のレコードを定義するステップ1120を、レコード読取り時間

507および更新時間610と併せ使用して、一次記憶制御装置405におけるアプリケーションWRITE入出力操作の正しい順序を決定することができる。

【0062】ステップ1160は、完全災害復旧の制約の下で特定の書き込み更新が二次DASD416に適用される、遠隔データ・シャドーイング・プロセスのバックエンドを表している。二次DASD416に更新内容を書き込む際に入出力エラーが発生するか、または二次側431全体が停止し、最初期設定された場合は、書き込みプロセスに入っていた整合性グループ全体を最初から再適用することができる。このため、どの二次DASD416の入出力が行われたか、どの入出力が行われなかったか、どの入出力が処理中かなどを追跡せずに、遠隔シャドーイングを行うことができる。

【0063】二次入出力書き込み

ステップ1160の重要な構成要素は、二次側431が一次側421から後れをとらないように、PDM414によってレコードが二次DASD416に効率よく書き込まれることである。必要な効率性、主に、様々な二次DASD416への複数の入出力操作を同時に実行することによって達成される。二次DASD416を一度に1つずつ連続して書き込むと、二次側431は一次側421からかなり遅れてしまう恐れがある。単一チャネル・コマンド・ワード(CCW)連鎖を介して単一の二次装置宛ての整合性グループごとにレコードを書き込めば、二次側431ではさらに高い効率性が得られる。それぞれの単一CCW連鎖内では、そこで行われる各二次DASD416のデータ・トラックへの入出力操作が一次ボリュームでの発生順に維持されている限り、その入出力操作をさらに最適化することができる。

【0064】特定の整合性グループ用の二次入出力操作を単一CCW連鎖内で最適化する場合、一部は一次書き込み入出力操作のパターンに基づいて行われ、一部は二次DASD416の物理特性に基づいて行われる。最適化は、二次DASD416がカウント/キー/データ(CKD)か、拡張カウント/キー/データ(ECKD)か、固定ブロック方式(FBA)かなどに応じて、多少変化する可能性がある。その結果、所与の時間間隔の間に1つの一次DASD406に対して行われる複数のWRITE入出力(m)は、1つの二次DASD416のボリュームに対する単一のSTART入出力操作に削減することができる。このように二次記憶制御装置415に対するSTART入出力の回数をm:1に最適化すると、二次DASD416は後れをとらずに済み、それにより、一次側421のレコード更新をもっと精密にシャドーイングすることができる。

【0065】正常な遠隔データ・シャドーイングとそれによる二次入出力の最適化の重要点は、一貫性のあるコピーを復旧に使用できるように、二次DASD416に対して同時に行う複数の入出力操作のいずれかで発生す

る回復不能エラーを最小限にすることである。所与の二次書き込みで失敗すると、その後の従属書き込みが条件付け書き込みを伴わずに記録される恐れがある(たとえば、実際にはデータベース用の実際の更新書き込みが失敗に終わっているのにデータベース・レコードが更新されたことを示すログ項目は、二次DASD416のコピーの順序整合性に違反する)。

【0066】その更新失敗が復旧されるまで、失敗した二次DASD416のコピーはアプリケーションの復旧に使用できなくなる。失敗した更新は、SDM414によってPDM404から現行コピーを要求することで修正できるはずである。その間に二次データ・コピーは不整合になり、そのため、PDM404が現行更新で応答し、それ以前の他のすべての更新がPDM404によって処理されるまで使用できなくなる。通常、失敗した更新の復旧に要する時間は、十分な災害復旧保護のために受け入れられないほど長い非復旧ウィンドウを示す。

【0067】効果的な二次側431の入出力最適化は、所与の整合性グループについて書き込まれるデータ・レコード・セットを検査し、ECKD方式などの二次DASD416の特定の方式の規則に基づいて連鎖を構築することで実現される。ここに開示する最適化技法は、入出力エラーが発生した場合に整合性グループを適用する際に、CCW連鎖を再実行できるように、または、二次初期プログラム・ロード(IPL)復旧の場合に、データを紛失せずに整合性グループ全体を再適用できるように、入出力エラーからの復旧を単純化するものである。

【0068】図12は、ECKD方式用のすべてのWRITE入出力の組合せに対応するCCW連鎖を構築するための完全整合性グループ復旧(FCGR)の規則を要約して示すもので、ここではCCHHRレコード形式が使用される(シリンダ番号、ヘッド番号、レコード番号)。図12は、1つの整合性グループの範囲内でDASDトラックに対して行われるWRITE入出力操作の可能な組合せをそれぞれ検査することによって作成される。図12のFCGR規則(図14および図15に記載する)は、整合性グループを適用する際のエラーについて完全復旧を行うためにデータ配置(二次DASD416の入出力書き込みCCW連鎖)を管理する場合に従うものである。図12に示すFCGR規則は、新しいWRITE入出力操作が追加されるたびに適切に拡張されることになる。これらの規則は、二次側431のハードウェアまたはソフトウェアで実現することができる。FCGR規則は、同一DASDトラック分析に対するREADレコード・セットを、一次DASD406のWRITE入出力タイプ、検索指数、およびカウント・フィールドとキー・フィールドの検査に還元するので都合である。

【0069】図12に示すように整合性グループの書き込み操作を検査せずにDASDトラックが書き込まれる

と、前に書き込まれたデータ・レコードの再書き込みができなくなる可能性がある。たとえば、以下の内容の連鎖があると想定する。

レコード5へのWRITE UPDATE

レコード1へのFORMAT WRITE

この場合、レコード1とレコード5は同じDASDトラックに存在し、レコード1がレコード5の前に置かれる。レコード5はUPDATE WRITE CCWによって更新されるが、FORMAT WRITE入出力CCWは、トラックの残りを消去してレコード1を更新するため、レコード5は削除される。この連鎖を再実行しなければならない場合、レコード5の先頭に配置するLOCATERECORD CCWが位置決めポイントを持たなくなる（レコード5が存在しなくなる）ため、この連鎖は先頭から完全に回復することができない。一次側421ですでに書き込み操作が正常に行われているので、データの整合性と保全性を維持するには、二次DASD416上の整合性グループ全体をいつでも適用することが必要である。

【0070】図16のステップ1410～1470は、図11のステップ1160によって表され、図12に定義されるFCGR規則を使用するプロセスの詳細を示すものである。ステップ1410でSDM414は、現行整合性グループの各種レコードを2つのカテゴリに分割する。第一のカテゴリは、同じ二次DASDボリューム向けの入出力命令を含み、第二のカテゴリは、第一のカテゴリに含まれるレコードのうち、同じCCHH向けのレコードの入出力命令を含む（すなわち、同じDASDトラックに更新されるレコード）。

【0071】現行整合性グループのレコードのカテゴリ分類後、ステップ1420では、トラック上のデータ配置を識別し、トラック／レコードのアドレス指定を行うために、アプリケーションのWRITE入出力およびSDM414のWRITE入出力を二次DASD416の方式、たとえば、ECKD方式のFCGR規則（図12を参照）に適合させる。SDM414は、ステップ1430で、同じボリュームに対する二次DASDのWRITE入出力操作を単一入出力CCW連鎖にグループ分けする。ステップ1440は、実際の二次DASD416の書き込み用の検索指数と特定のレコード・データ（CKDフィールド）に応じて、それぞれの二次DASD416のヘッド・ディスク・アセンブリ（HDA）を移動させることを含む。

【0072】ステップ1450では、後続の書き込み操作によって、前の書き込み操作またはDASD検索指数（ここで消去されるレコードでの分割など）を無効にするかどうかを判別するために図12のFCGR規則を使用して、第二のカテゴリ（通常、レコードを受け取るトラックごとに1つずつ、複数の第二のカテゴリが存在する）を構成するこれらのレコードについてREAD SET

BUFFERS1と2を比較する。READ SET BUFFERS1と2は隣接する読取りレコード・セットが含まれている。

FCGR規則に従うと、エラーが発生した場合に、一次側421からレコード更新を再度受け取らなくても、SDM414は整合性グループ全体を再書き込みできるようになる。SDM414が現行整合性グループを二次DASD416に適用した後、ステップ1460では、状態テーブル（図7）とマスタ・ジャーナル（図8）を更新する。

【0073】ステップ1470は、次の整合性グループ（現行整合性グループになるもの）を獲得して、処理をステップ1410に戻すので、遠隔コピー・プロセスはリアルタイムで続行される。一次側421から二次側431への通信が終了した場合、遠隔コピー・プロセスは停止する。この通信は、ボリューム対がPDM404によってプロセスから削除された場合、一次側が破壊された場合（災害が発生した場合）、規則的な運転停止が行われた場合、または二次側431で特定の引継ぎ処置が行われた場合に終了することがある。二次側431でジャーナル処理された整合性グループは、引継ぎ操作中に二次DASD416に適用することができる。一次側421が捕捉したデータのうち、SDM414によって完全に受け取られていないデータだけが紛失する。

【0074】要約すると、これまで同期および非同期遠隔データ二重化システムについて説明してきた。非同期遠隔データ二重化システムは、記憶域ベースのリアルタイム・データ・シャドーイングを提供する。一次側は、レコード更新を生成するアプリケーションを実行し、一次側から離れた位置にある二次側は、レコード更新をシャドーイングして、一次側の災害復旧を行う。この非同期遠隔データ二重化システムは、一次側の時間依存プロセスを同期させるためのシスブックス・クロックと、アプリケーションを実行するための一次側の一次プロセッサとを含み、一次プロセッサはそこに一次データ・ムーバを有している。一次プロセッサには、各レコード更新ごとに書き込み入出力操作を出すために複数の一次記憶制御装置が連結され、それぞれの一次記憶制御装置DASD書き込み入出力操作がシスブックス・クロックに同期している。複数の一次記憶装置はこの書き込み入出力操作を受け取り、それに応じてレコード更新をそこに格納する。一次データ・ムーバは、各レコード更新ごとに複数の一次記憶制御装置からレコード・セット情報を収集し、所定のグループのレコード・セット情報に接頭部ヘッダを付加する。この接頭部ヘッダと所定のグループのレコード・セット情報が自己記述レコード・セットを形成する。それぞれのレコード・セット情報は、一次装置アドレス、シリンダ番号とヘッド番号（CCHH）、レコード更新順序番号、書き込み入出力タイプ、検索指数、セクタ番号、およびレコード更新時間を含む。接頭部ヘッダは、総データ長、操作タイム・スタンプ、時間間隔

グループ番号、およびレコード読取り時間を含む。二次側の二次プロセッサは二次データ・ムーバを有し、この二次データ・ムーバが一次側から自己記述レコード・セットを受け取る。二次プロセッサには複数の二次記憶制御装置が連結され、二次記憶制御装置にはレコード・更新のコピーを格納するために複数の二次記憶装置が連結されている。二次データ・ムーバは、送られてきた自己記述レコード・セットが完全なものであるかどうかを判定し、その自己記述レコード・セットから整合性グループを形成し、さらに、レコード更新が複数の一次記憶装置に書き込まれた順序と整合する順序で複数の二次記憶装置に書き込むために、各整合性グループから得たレコード更新を複数の二次記憶制御装置に出力する。

【0075】特に本発明の実施例に関連して本発明を図示し説明してきたが、当業者は、本発明の精神および範囲を逸脱せずに形態および詳細の様々な変更が可能であることに留意されたい。たとえば、整合性グループは、受け取った自己記述レコード・セットに基づいて二次データ・ムーバによって形成されたものとして説明してきたが、書き込みレコード・セットに基づいて一次側で整合性グループを形成するか、二次側の他の部分で形成することもできる。一次側と二次側の記憶装置の形式は、同じである必要はない。たとえば、CKDレコードを固定ブロック方式(FBA)タイプのレコードなどに変換することも可能である。また、記憶装置は、DASD装置に限定されているわけではない。

【0076】

【発明の効果】本発明の実施により、災害復旧のためにDASDデータを二次側にシャドーイングするための改良された設計および方法を提供することができる。

【0077】まとめとして、本発明の構成に関して以下の事項を開示する。

【0078】(1) 一次プロセッサで実行される1つまたは複数のアプリケーションによって生成されたデータ更新が一次記憶サブシステムによって受け取られ、一次記憶サブシステムは入出力書き込み操作によって各データ更新を書き込まれ、各書き込み入出力操作にタイム・スタンプが付けられ、共通タイムによってタイム・スタンプの同期が取られ、一次プロセッサと同じ地域にあるか一次プロセッサから離れた位置にあるかにかかわらず、二次システムは、災害復旧のために二次側が使用できるように順序整合性のある順序でデータ更新をシャドーイングする、災害復旧機能を提供するために整合性グループを形成する方法において、前記方法が、(a) 一次記憶サブシステムで発生する各書き込み入出力操作にタイム・スタンプを付けるステップと、(b) 各データ更新ごとに一次記憶サブシステムから書き込み入出力操作レコード・セット情報を収集するステップと、(c) データ更新とそれぞれのレコード・セット情報から自己記述レコード・セットを生成し、この自己記述レコード・セット

が、二次システムだけによる書き込み入出力操作の順序を再生成するために十分な制御情報を含んでいるステップと、(d) 自己記述レコード・セットを間隔グループにグループ分けし、各間隔グループが、操作タイム・スタンプ開始時間から測定され、所定の間隔しきい値の間持続するステップと、(e) 最も早い操作タイム・スタンプを有する自己記述レコード・セットの間隔グループとして現行整合性グループを選択し、一次記憶サブシステムでの入出力書き込み操作の時間順に基づいて、個々のデータ更新が現行整合性グループ内で順序付けされるステップとを含む方法。

(2) 各間隔グループの開始時間を識別するための操作タイム・スタンプに基づいて、一次記憶サブシステムとのセッションを開始し、各間隔グループの境界が連続する操作タイム・スタンプによって示されることが、ステップ(b)にさらに含まれることを特徴とする、上記(1)に記載の方法。

(3) 各間隔グループを記述する接頭部ヘッダを追加することがステップ(d)に含まれることを特徴とする、上記(1)に記載の方法。

(4) 自己記述レコード・セットからなる間隔グループを二次側に送信するステップ(f)をさらに含むことを特徴とする、上記(3)に記載の方法。

(5) 受け取った各自己記述レコード・セットが完全なものであるかどうかを二次側で判定するステップ(g)をさらに含むことを特徴とする、上記(4)に記載の方法。

(6) 自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落データ更新を再送信するよう二次側が一次側に要求することが、ステップ(g)にさらに含まれることを特徴とする、上記(5)に記載の方法。

(7) 各時間間隔グループが完全なものであるかどうかを二次側で判定するステップ(h)をさらに含むことを特徴とする、上記(6)に記載の方法。

(8) 間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう二次側が一次側に要求することが、ステップ(h)にさらに含まれることを特徴とする、上記(7)に記載の方法。

(9) 整合性グループで順序づけされたように、対応する一次側の書き込み入出力操作の順序に応じて、二次側で受け取ったデータ更新を二次記憶サブシステムに書き込むステップ(i)をさらに含むことを特徴とする、上記(8)に記載の方法。

(10) 災害復旧のために遠隔データ・シャドーイングを行うシステムにおいて、このシステムは、一次データ・ムーバと、レコード更新を生成するアプリケーションとを実行する一次プロセッサを有する一次側を含み、一次プロセッサは、一次プロセッサから一次記憶サブシステムに出される書き込み入出力操作に応じてレコード更新

を格納するための記憶装置を有する一次記憶サブシステムに連結され、一次側は、一次側の時間依存操作を同期させるために共通のシステム・タイマをさらに含み、システムは、一次プロセッサと通信する二次プロセッサと、順序整合性のある順序でレコード更新のコピーを格納するための二次記憶サブシステムとを有する二次側をさらに含み、(a) 一次記憶サブシステムの各書き込み入出力操作にタイム・スタンプを付けるステップと、

(b) 一次記憶サブシステム内の各記憶装置とのセッションを確立するステップと、(c) 一次記憶サブシステム内の各記憶装置からレコード・セット情報を収集するステップと、(d) レコード・セットとそれぞれのレコード・セット情報を一次データ・ムーバに読み込むステップと、(e) 各レコード・セットの前にヘッダを付け、それに基づく自己記述レコード・セットを生成するステップと、(f) 所定の時間間隔に応じて、時間間隔グループ単位で自己記述レコード・セットを二次プロセッサに送信するステップと、(g) 自己記述レコード・セットから整合性グループを形成するステップと、

(h) 順序整合性のある順序で各整合性グループのレコード更新を二次記憶サブシステムにシャドーイングするステップとを含む、順序整合性のある順序でレコード更新をシャドーイングするための方法。

(11) レコード・セットが非同期的に二次プロセッサに送信されることを特徴とする、上記(10)に記載の方法。

(12) ステップ(g)が二次側で行われることを特徴とする、上記(10)に記載の方法。

(13) 受け取った各自己記述レコード・セットが完全であるかどうかを二次側で判定するステップが、ステップ(f)にさらに含まれることを特徴とする、上記(10)に記載の方法。

(14) 受け取った自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落レコード更新を再送信するよう一次側に要求するステップが、ステップ(f)にさらに含まれることを特徴とする、上記(13)に記載の方法。

(15) 各時間間隔グループが完全であるかどうかを二次側で判定するステップ(i)をさらに含むことを特徴とする、上記(10)に記載の方法。

(16) 時間間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう一次側に要求することが、ステップ(i)にさらに含まれることを特徴とする、上記(15)に記載の方法。

(17) ステップ(c)が、レコード・セット情報において、各レコード更新が格納されている一次記憶装置上の物理的位置を識別することを特徴とする、上記(10)に記載の方法。

(18) ステップ(c)が、レコード・セット情報において、セッション内に一次記憶装置に格納された各レコ

ード更新の順序と更新時間を識別することを特徴とする、上記(17)に記載の方法。

(19) ステップ(e)が、接頭部ヘッダにおいて、セッション用の間隔グループ番号と、そこで参照される各レコード更新用のグループ内順序を識別することを特徴とする、上記(10)に記載の方法。

(20) 1つまたは複数のアプリケーションを実行する一次プロセッサを有し、この1つまたは複数のアプリケーションがレコード更新を生成し、一次プロセッサがそれに基づく自己記述レコード・セットを生成し、自己記述レコード・セットが二次システムに送られ、二次システムがリアルタイム災害復旧のために自己記述レコード・セットに基づいて順序整合性のある順序でレコード更新をシャドーイングし、一次プロセッサが一次記憶サブシステムに連結され、一次記憶サブシステムがレコード更新を受け取って、そこに各レコード更新を格納するために書き込み入出力操作を実行し、一次プロセッサが、同期を取るためにアプリケーションと一次記憶サブシステムに共通の時間源を提供するためのシブプレックス・クロックと、各レコード更新ごとにレコード・セット情報を提供するように一次記憶サブシステムに指示し、複数のレコード更新と、それに対応するそれぞれのレコード・セット情報を時間間隔グループにグループ分けし、それに接頭部ヘッダを挿入し、各時間間隔グループがレコード・セットを自己記述する一次データ・ムーバ手段とを含む一次システム。

(21) 一次記憶サブシステムが、書き込み入出力操作を出す複数の一次記憶制御装置と、複数の一次記憶サブシステムに連結された複数の一次記憶装置とを含むことを特徴とする、上記(20)に記載の一次システム。

(22) 複数の一次記憶装置が直接アクセス記憶装置であることを特徴とする、上記(21)に記載の一次システム。

(23) 一次データ・ムーバ手段が、各時間間隔グループに關与する複数の一次記憶制御装置のうちの各一次記憶制御装置用のそれぞれの書き込み入出力操作ごとにレコード・セット情報を収集することを特徴とする、上記(22)に記載の一次システム。

(24) 一次プロセッサにおいて、各書き込み入出力操作にシブプレックス・クロックに關するタイム・スタンプが付けられ、各書き込み入出力操作が、複数の一次記憶制御装置のうちの1つの一次記憶制御装置に出され、各一次記憶制御装置が、タイム・スタンプを保持し、対応する読取りレコード・セットに入れてそのタイム・スタンプを一次データ・ムーバ手段に返すことを特徴とする、上記(23)に記載の一次システム。

(25) 各レコード・セット情報が、複数の一次記憶装置のうちの1つの一次記憶装置上における対応するレコード更新の物理的位置を識別することを特徴とする、上記(24)に記載の一次システム。

(26) 各レコード・セット情報が、対応するレコード更新の一次サブシステムID、一次装置アドレス、シリンダ番号、およびヘッド番号を識別することを特徴とする、上記(24)に記載の一次システム。

(27) 一次データ・ムーバ手段が、1つの時間間隔グループに關与するすべての一次記憶制御装置にわたる各書込み入出力更新の相対順序を識別することを特徴とする、上記(24)に記載の一次システム。

(28) レコード更新をジャーナル処理し、一次システムおよび二次システム上の各レコード更新の記憶場所を相互参照するために、一次データ・ムーバ手段が状態テーブルを作成することを特徴とする、上記(27)に記載の一次システム。

(29) 一次データ・ムーバ手段が、二次システムに状態テーブルを送信することを特徴とする、上記(27)に記載の一次システム。

(30) 一次側と二次側を含み、二次側が災害復旧のために一次側のレコード更新をリアルタイムでシャドーイングし、レコード更新が一次側で実行されるアプリケーションによって生成され、一次側が、シスプレックス・クロックと、レコード更新を生成するアプリケーションを実行し、各レコード更新ごとに対応する書込み入出力操作を出し、一次データ・ムーバをそこに有する一次プロセッサと、レコード更新を格納するよう指示され、各レコード更新ごとに出された書込み入出力操作を実行する複数の一次記憶制御装置と、対応する書込み入出力操作に応じて、レコード更新を受け取ってそこに格納する複数の一次記憶装置とを含み、書込み入出力操作が互いに正しい順序で並べられるように、シスプレックス・クロックによって同期が取られた通りに、一次プロセッサによって一次プロセッサと各書込み入出力にタイム・スタンプが付けられ、一次データ・ムーバが、複数組のレコード更新を収集し、複数の一次記憶制御装置のうちのそれぞれによって提供された各レコード・セット情報と対応するレコード更新を組み合わせ、各レコード・セット情報が、それぞれの対応する書込み入出力操作の相対順序と時間を含み、一次データ・ムーバが時間間隔グループ別にレコード更新を収集して、各時間間隔グループに接頭部ヘッダを挿入し、接頭部ヘッダが各時間間隔グループに含まれるレコード更新を識別する情報を含み、各レコード・セット情報と接頭部ヘッダが、自己記述レコード・セットを生成するために組み合わせられ、自己記述レコード・セットが二次側に送信され、自己記述レコード・セットが、一次側からの追加の通信がなくても順序整合性のある順序でレコード更新をそこにシャドーイングするために十分な情報を二次側に提供する、遠隔データ・シャドーイング・システム。

(31) 遠隔データ・シャドーイングに關与するために識別されたすべての一次記憶制御装置とのセッションを確立することによって、一次データ・ムーバが時間間隔

グループを形成することを特徴とする、上記(30)に記載の遠隔データ・シャドーイング・システム。

(32) 一次データ・ムーバが、識別されたすべての一次記憶制御装置からレコード・セット情報を収集することを特徴とする、上記(31)に記載の遠隔データ・シャドーイング・システム。

(33) 一次プロセッサが、二次側に自己記述レコードを送信することを特徴とする、上記(32)に記載の遠隔データ・シャドーイング・システム。

(34) 一次データ・ムーバが、自己記述レコードから整合性グループを形成することを特徴とする、上記(32)に記載の遠隔データ・シャドーイング・システム。

(35) 二次側が、送信された自己記述レコードから整合性グループを形成することを特徴とする、上記(33)に記載の遠隔データ・シャドーイング・システム。

(36) レコード更新をジャーナル処理し、一次システムおよび二次システム上の各レコード更新の記憶場所を相互参照する状態テーブルを、一次データ・ムーバが作成することを特徴とする、上記(32)に記載の遠隔データ・シャドーイング・システム。

(37) 複数の一次記憶装置が直接アクセス記憶装置(DASD)であることを特徴とする、上記(32)に記載の遠隔データ・シャドーイング・システム。

(38) 各レコード・セット情報が、一次装置アドレスと、シリンダ番号およびヘッド番号(CCHH)と、レコード更新順序番号と、書込み入出力タイプと、検索指数と、セクタ番号と、レコード更新時間とを含むことを特徴とする、上記(37)に記載の遠隔データ・シャドーイング・システム。

(39) 接頭部ヘッダが、総データ長と、操作タイム・スタンプと、時間間隔グループ番号と、レコード読取り時間とを含むことを特徴とする、上記(37)に記載の遠隔データ・シャドーイング・システム。

(40) レコード更新を生成するアプリケーションを実行する一次側を含み、一次側から離れた位置に二次側を有し、二次側がレコード更新をシャドーイングして、一次側に災害復旧を提供する、記憶域ベースのリアルタイム・データ・シャドーイングを行う非同期遠隔データ二重化システムにおいて、非同期遠隔データ二重化システムが、一次側の時間依存プロセスを同期させるためのシスプレックス・クロックと、アプリケーションを実行し、対応するレコード更新用の書込み入出力操作を出し、一次データ・ムーバをそこに有する、一次側の一次プロセッサと、各レコード更新ごとに書込み入出力操作を1つずつ受け取り、それぞれの一次記憶制御装置書込み入出力操作が一次プロセッサによってシスプレックス・クロックと同期している、複数の一次記憶制御装置と、対応する書込み入出力操作に応じて、レコード更新をそこに格納するための複数の一次記憶装置とを含み、一次データ・ムーバが、各レコード更新ごとに複数の一

次記憶制御装置からレコード・セット情報を収集して、
所定のグループのレコード・セット情報に接頭部ヘッダ
を付加し、接頭部ヘッダと所定のレコード・セット情報
グループが自己記述レコード・セットを形成し、各レコ
ード・セット情報が、一次装置アドレス、シリンダ番号
およびヘッド番号（CCHH）、レコード更新順序番
号、書き込み出力タイプ、検索指数、セクタ番号、およ
びレコード更新番号を含み、接頭部ヘッダが、総データ
長、操作タイム・スタンプ、時間間隔グループ番号、お
よびレコード読取り時間を含み、二次データ・ムーバを
有し、その二次データ・ムーバが一次側から自己記述レ
コード・セットを受け取る、二次側の二次プロセッサ
と、二次プロセッサに連結された複数の二次記憶制御装
置と、レコード更新を格納する複数の二次記憶装置とを
さらに含み、二次データ・ムーバが、送信された自己記
述レコード・セットが完全なものであるかどうかを判定
し、自己記述レコード・セットから整合性グループを形
成し、複数の一次記憶装置にレコード更新が書き込まれ
たときの順序に整合する順序で複数の二次記憶装置に書
き込むために各整合性グループから得たレコード更新を
複数の二次記憶制御装置に出力する、非同期遠隔データ
二重化システム。

【図面の簡単な説明】

【図1】同期遠隔データ・シャドーイング機能を有する
災害復旧システムのブロック図である。

【図2】図1の災害復旧システムにより同期遠隔コピー
を提供する方法を示す流れ図である。

【図3】入出力エラー回復プログラム（入出力ERP）
操作の方法を示す流れ図である。

【図4】非同期遠隔データ・シャドーイング機能を有す
る災害復旧システムのブロック図である。

【図5】図4の一次側からの読取りレコード・セットの
前に付く接頭部ヘッダを示すデータ形式図である。

【図6】読取りレコード・セットを構成する各種フィー
ルドを示すデータ形式図である。

【図7】ボリューム構成情報を識別する状態テーブルで
ある。

【図8】図4の二次側が使用するマスタ・ジャーナルで

ある。

【図9】整合性グループを形成するためのシーケンス例
である。

【図10】整合性グループを形成するために情報および
読取りレコード・セットを収集する方法を示す流れ図で
ある。

【図11】整合性グループを形成する方法を示す流れ図
である。

【図12】DASDトラックに対する所与の入出力操作
シーケンスの場合のECKD方式装置用の完全整合性グ
ループ復旧規則アプリケーションを示すテーブルであ
る。

【図13】図12のテーブルで使用する規則の説明の構
成を示す図である。

【図14】図12のテーブルで使用する規則の説明の一
部である。

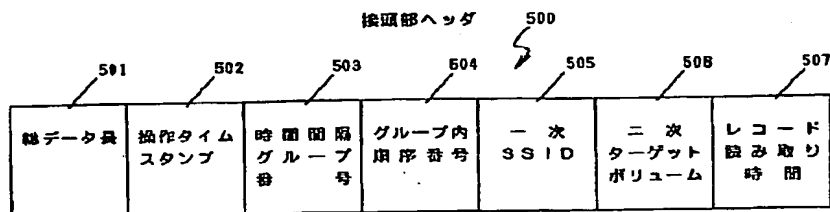
【図15】図12のテーブルで使用する規則の説明の一
部である。

【図16】完全整合性グループ復旧機能を持つ二次側に
読取りレコード・セット・コピーを書き込む方法の流れ
図である。

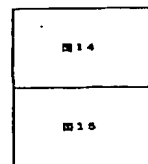
【符号の説明】

- 1 一次プロセッサ
- 2 入出力ERP
- 3 一次記憶制御装置
- 4 一次DASD
- 5 二次プロセッサ
- 6 二次記憶制御装置
- 7 二次DASD
- 8 対等通信リンク
- 9 エンタープライズ・システム接続（ESCON）リ
ンク
- 10 災害復旧システム
- 11 ホスト間通信リンク
- 12 チャンネル
- 13 チャンネル
- 14 一次側
- 15 二次側

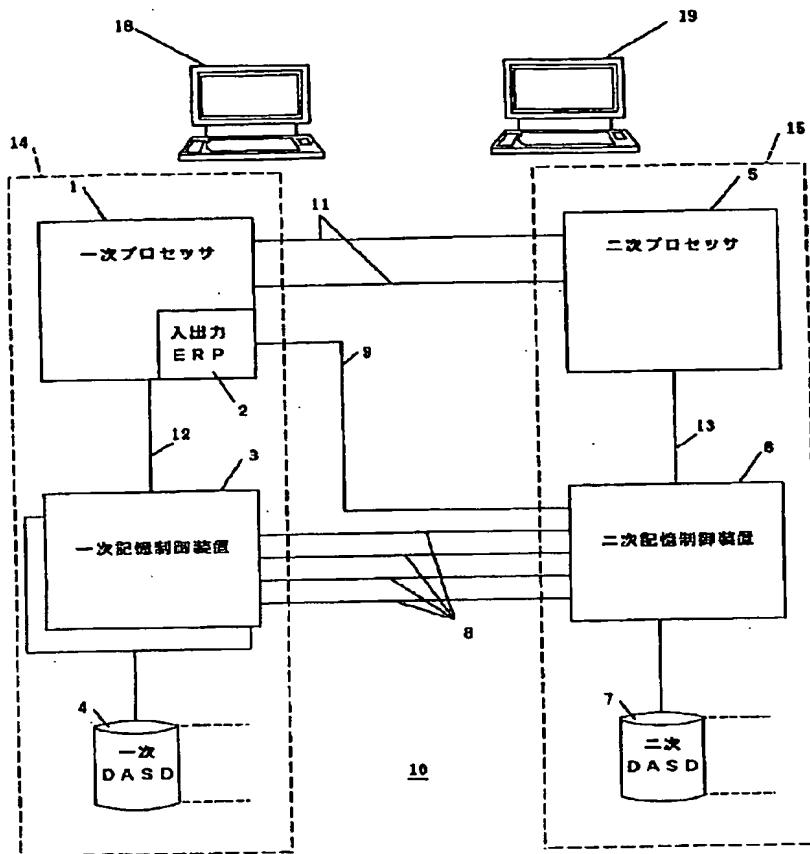
【図5】



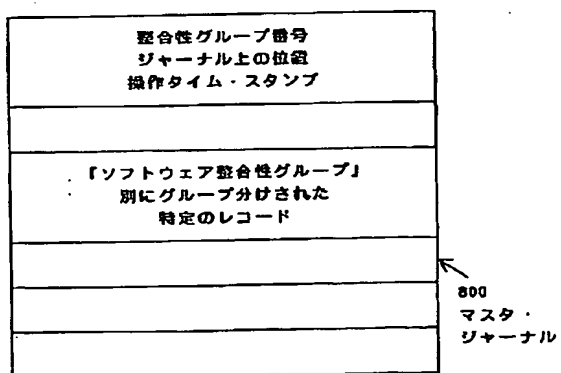
【図13】



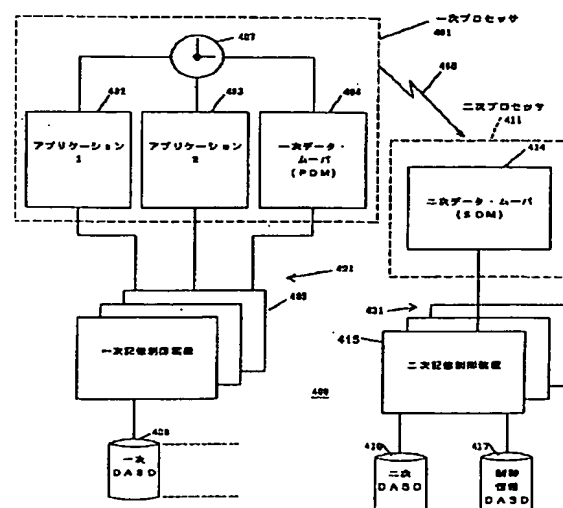
【図1】



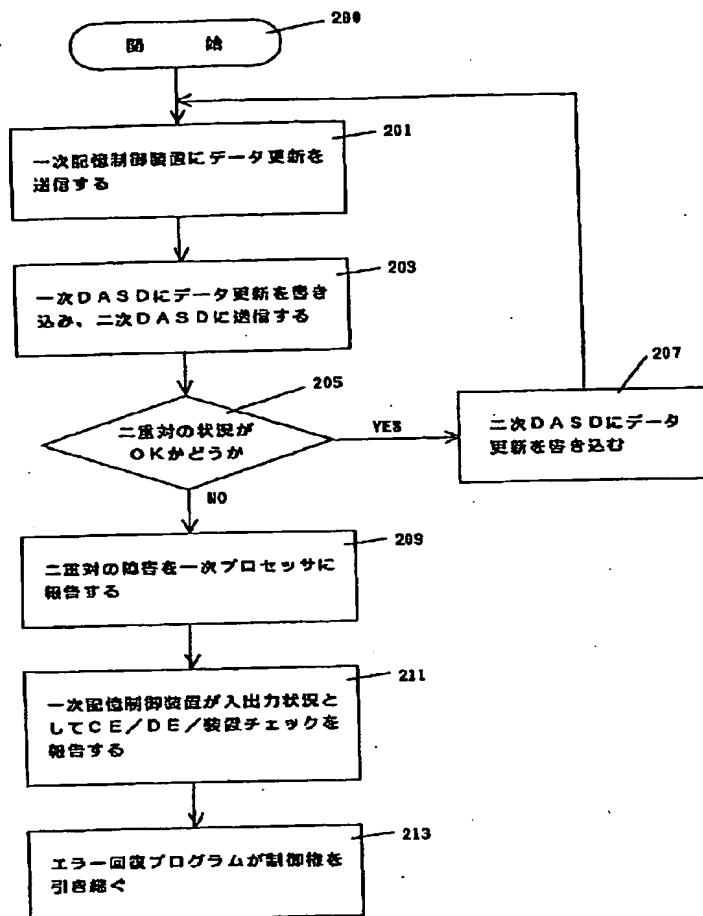
【図8】



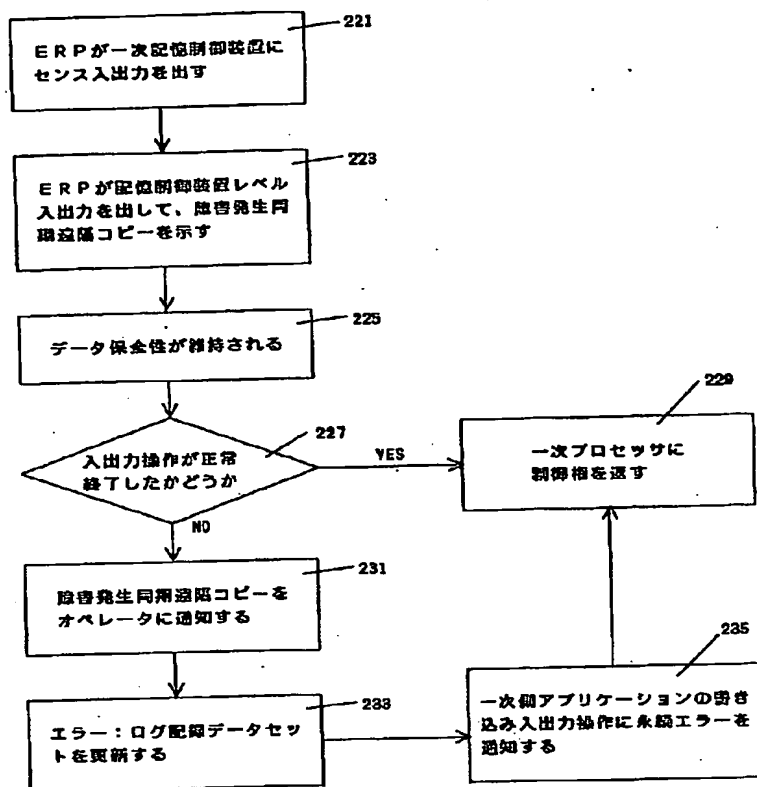
【図4】



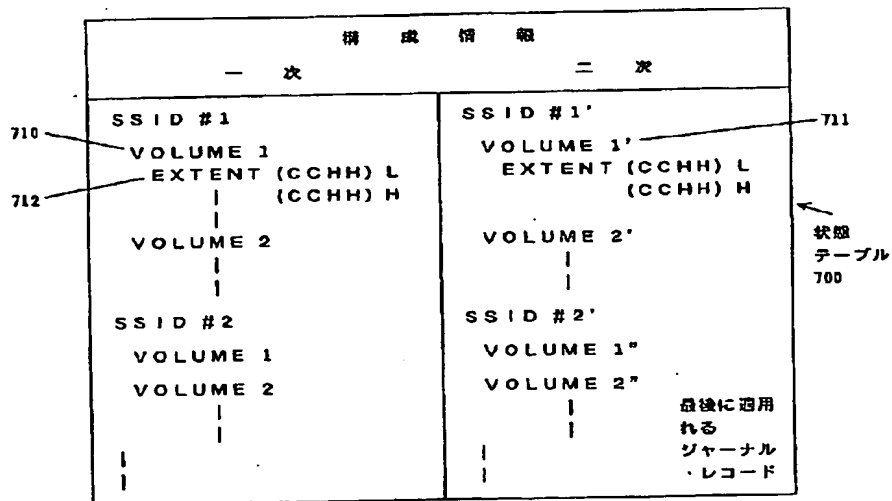
【図2】



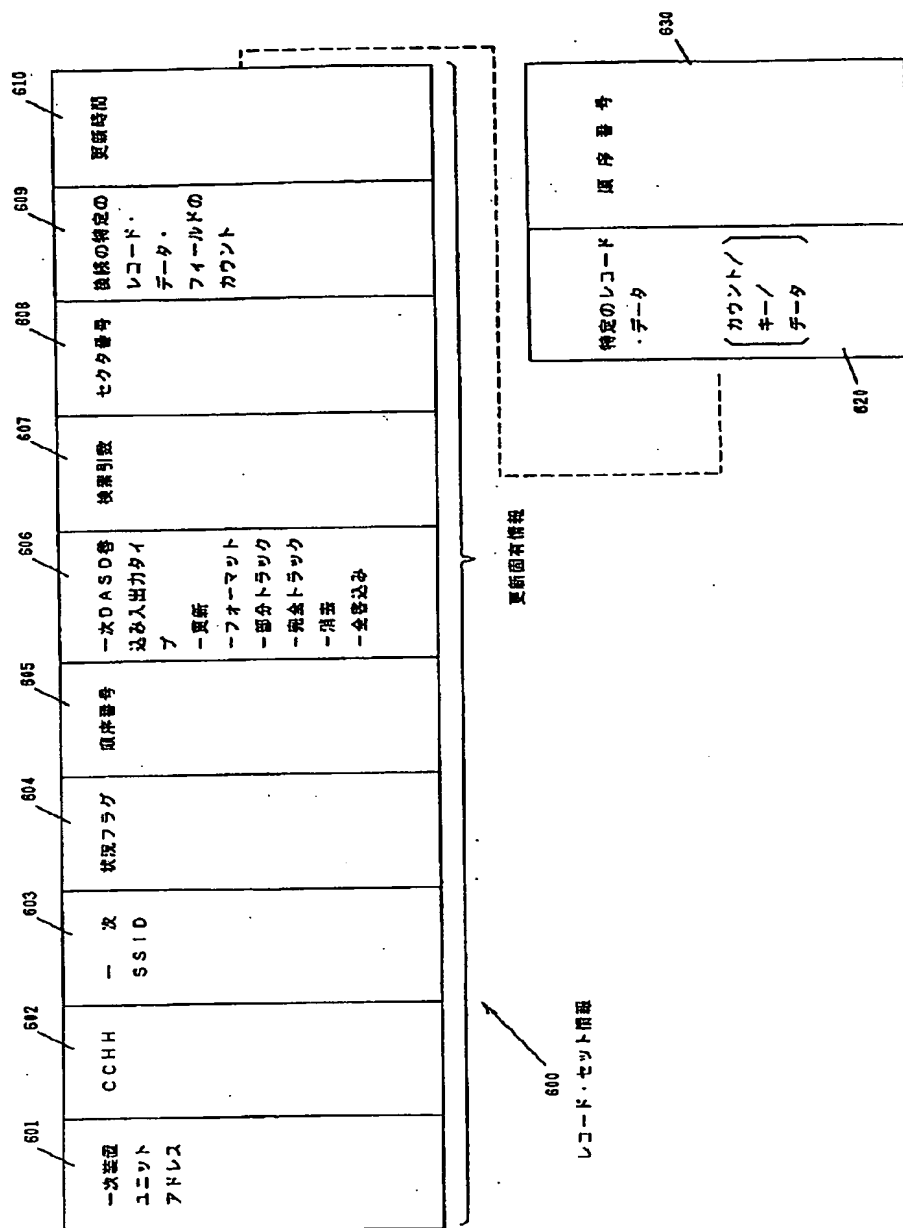
【図3】



【図7】



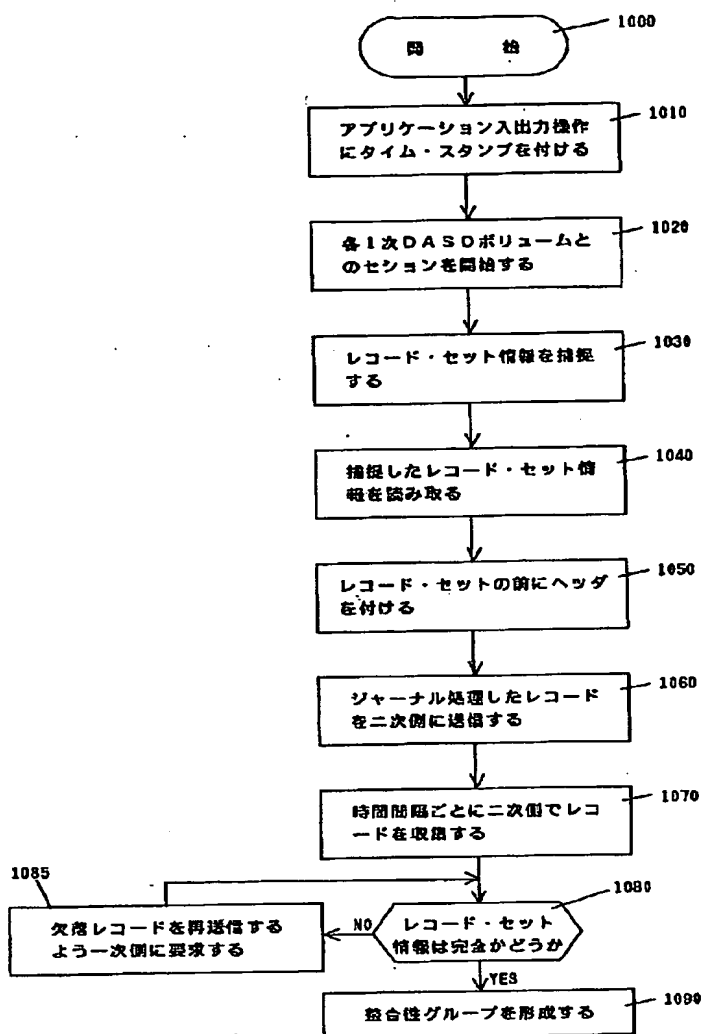
【図6】



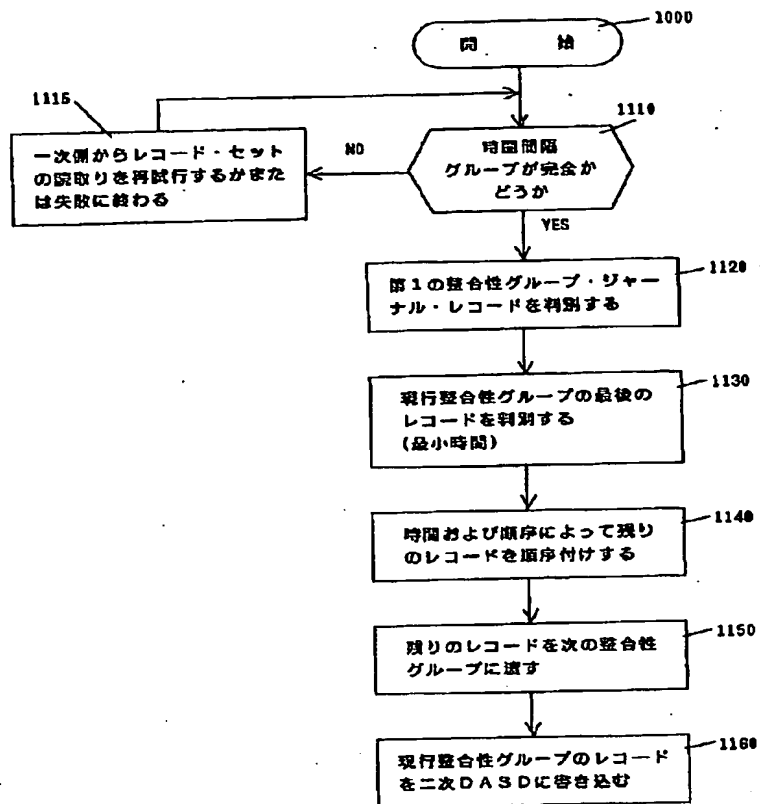
【図9】

物理制御 装置ID	操作タイム ・スタンプ	時間間隔 グループ 番号	読み取りレコード・セット 更新時間/制御装置		
			順序3の1	順序3の2	順序3の3
SSID1	T1	G1	11:59 ②	12:00 ⑤	12:01 ⑧
SSID2	T1	G1	12:00 ④	12:02 ⑦	
SSID3	T1	G1	11:58 ①	11:59 ③	12:02 ⑥
SSID1	T2	G2			
SSID2	T2	G2			
SSID3	T2	G2			
SSID3	T3	G3			
<u>整合性グループ番号1</u>					
① 11:58					
② 11:59					
③ 11:59					
④ 12:00					
⑤ 12:00					
⑥ 12:01					
読み取り順に並べた場合の最も早い操作時間T1 SSID全体で最も早い更新時間					
SSID全体での最大更新時間の最も早い時間					
<u>整合性グループ番号2</u>					
⑦ -----					
⑧ -----					

【図10】



【図11】



【図12】

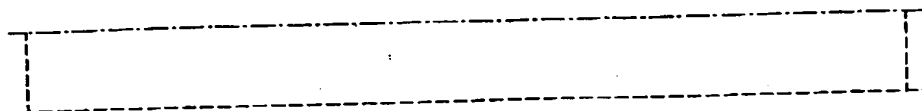
完全整合性グループ回復規則

既取リレコード・ セット・パツファ番号2		既取リレコード・セット・パツファ番号1									
入出力帯込み操作のタイプ											
	帯込み更新 KL=0	帯込み更新 KL≠0	完全フォア マツト帯込み	部分フォア マツト帯込み	完全除去	部分除去	任意帯込み KL=0	任意帯込み KL≠0			
帯込み更新 KL=0	W*	E*	N	J	D	K	W	E*			
帯込み更新 KL≠0	E*	W*	N	J	D	K	E*	W			
完全フォアマツト帯込み	T	T	R	T	R	T	R	R			
部分フォアマツト帯込み	C	C	N	H	P	L	W	W			
完全除去	T	T	R	R	R	T	T	T			
部分除去	B	B	N	M	E	G	W	W			
任意帯込み KL=0	W	E*	W	W	E	W	W	E*			
任意帯込み KL≠0	E*	W	W	W	E	W	E*	W			

【図14】

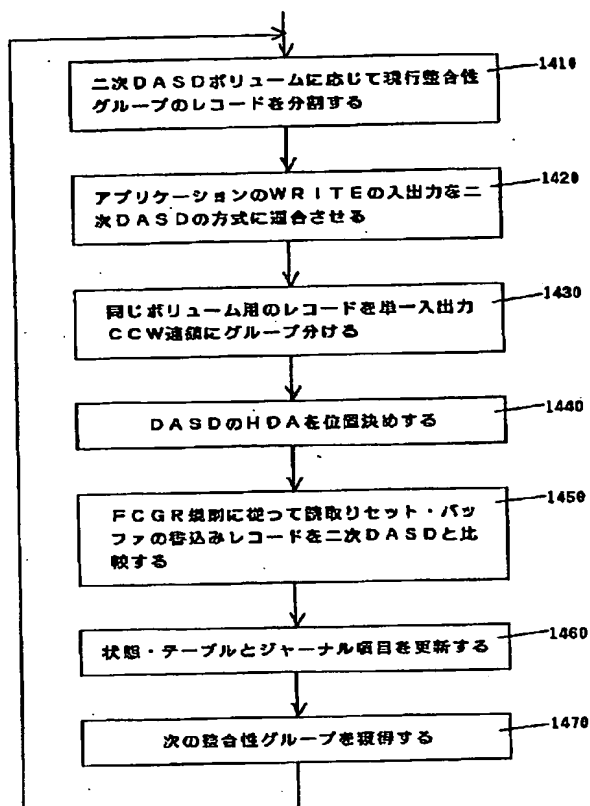
- B -番号1の検索が番号2の検索回数より大きい場合、番号1を捨てるが両方とも実行する
 - C -番号1のレコードが番号2に第1のレコードと等しいがそれ以上である場合、番号1を捨てるが両方とも実行する
 - D -番号2がR0を更新している場合、両方とも実行するがエラーになる
 - E -エラー（発生してはならない）
 - E* -番号1と番号2が同じレコードである場合、エラーになる
（両者間で形式書き込みを行わずに発生してはならない）
 - F -番号2の第1のレコードがR1である場合、両方とも書き込むが、エラーになる
 - G -番号1の検索回数が番号2の検索回数と等しいがそれ以上である場合、番号1を捨てるがエラーになる
 - H -番号1の検索回数が番号2の最後のレコードより大きい場合、番号1を捨てる。または番号2の検索回数が番号1の最後のレコードより大きい場合、エラーになるが両方とも書き込む
-
- X -さらに最適化するには以下の手順を実行できる
（番号1の検索回数が番号2の最後のレコードと等しいがそれ以上である）または（番号1の最後のレコードが番号2の最後のレコードと等しいがそれ以上である）しかも（番号2の検索回数が番号1の最後のレコードと等しいが、それ以下である）場合番号1を捨てる
または（番号2の検索回数が番号1の最後のレコードより大きい）
場合
エラーになる
または両方とも書き込む

【図15】



- J - 番号2のレコード（または検索引数）が番号1の最後のレコードより大きい場合、エラーになるが両方とも書き込む
- K - 番号2のレコード（または検索引数）が番号1の検索引数より大きい場合、エラーになるが両方とも書き込む
- L - 番号1の検索引数が番号2の検索引数と等しいかそれ以上である場合、両方とも書き込むがまたは番号1を捨てるかあるいはエラーになる
- M - （番号1の検索引数が番号2の検索引数と等しいかそれ以上である）場合、番号1を捨てるまたは両方とも書き込む
- N - 番号2の検索引数が番号1の最後のレコードより大きい場合、エラーになるが両方とも書き込む
- R - 番号1を捨てるもよい
- T - 番号1を捨てるなければならない
- W - 両方とも書き込む
- W* - 番号1と番号2が同じレコードを持つ場合、番号1を捨てるかまたは両方とも実行するかまたはレコードを組み合わせて一方の書き込みを実行する

【図16】



フロントページの続き

(72)発明者	ロナルド・マイナード・カーン アメリカ合衆国85748 アリゾナ州ツー ソン ノース・コレット・プレイス 35 761	(56)参考文献	特開 平7-6099 (JP, A) 特開 平5-204739 (JP, A) 特開 平4-33027 (JP, A) 特開 昭55-87262 (JP, A) 特開 昭64-67675 (JP, A) 特開 昭63-138441 (JP, A) 特開 平5-334161 (JP, A)
(72)発明者	グレゴリー・エドワード・マックブライ ド アメリカ合衆国85715 アリゾナ州ツー ソン イースト・フェアマウント・プレ イス 8622 40		
(72)発明者	デヴィッド・マイケル・シャクルフォ ード アメリカ合衆国85705 アリゾナ州ツー ソン ウェスト・サーバー・プレイス・ ドライブ 1348 45	(58)調査した分野(Int.Cl. ⁷ , DB名)	G06F 3/06 G06F 11/16 - 11/20 G06F 12/00, 12/16

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

EP 1 049 016 A2

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
02.11.2000 Bulletin 2000/44(51) Int. Cl.⁷: G06F 11/14

(21) Application number: 00103718.3

(22) Date of filing: 22.02.2000

(84) Designated Contracting States:

AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE

Designated Extension States:

AL LT LV MK RO SI

(30) Priority: 26.04.1999 JP 11767099

(71) Applicant: Hitachi, Ltd.
Chiyoda-ku, Tokyo 101-8010 (JP)

(72) Inventors:

- Tabuchi, Hideo,
Hitachi, Ltd. Intelle. Prop. Group
Chiyoda-ku, Tokyo 100-8220 (JP)

- Nozawa, Masafumi,
Hitachi, Ltd. Intel. Prop. Group
Chiyoda-ku, Tokyo 100-8220 (JP)
- Shimada, Akinobu,
Hitachi, Ltd. Intel. Prop. Group
Chiyoda-ku, Tokyo 100-8220 (JP)

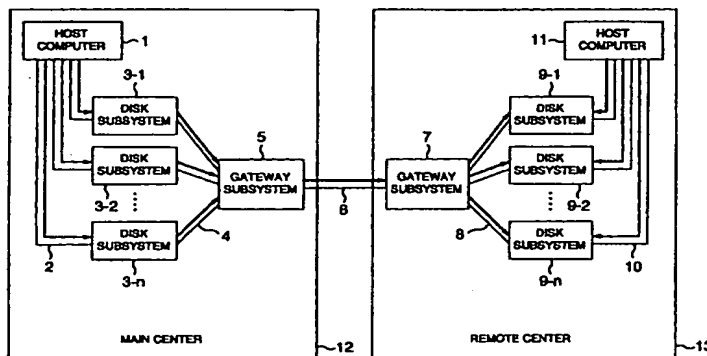
(74) Representative:
Strehl Schübel-Hopf & Partner
Maximilianstrasse 54
80538 München (DE)

(54) Disk subsystems and their integrated system

(57) An asynchronous remote copy system is provided which can ensure the data renewal order and data integrity of the disk subsystems and are easy to be incorporated and free from degradation of the process performance of host computers (1, 11). To this end, in the remote copy system for data mirroring, a main center (12) has one gateway subsystem (5) and a remote center (13) has one gateway subsystem (7), and disk subsystems (3, 9) in each center to be remotely copied are connected to the corresponding gateway subsystem. Data is mirrored through synchronous type

remote copy between a volume of the disk subsystem of each center to be remotely copied and a desired volume of the corresponding gateway subsystem, and the gateway subsystem of the main center sends the renewal data to the gateway subsystem of the remote center in accordance with the order of renewal of volumes of the gateway subsystem of the main center, to make the gateway subsystem of the remote center reflect the renewal data upon the volumes thereof through asynchronous type remote copy.

FIG.1



EP 1 049 016 A2

Description

BACKGROUND OF THE INVENTION

[0001] The present invention relates to external storage devices for storing data in a computer system and to their integrated system, and more particularly to remote copy techniques for mirroring data between remote external storage devices (disk subsystems) without involving an upper hierarchical apparatus or host computer, by interconnecting remote external storage devices and other remote external storage devices. The disk subsystem is herein intended to mean a control unit for controlling data transfer to and from an upper hierarchical apparatus and a storage device having disks for storing data or a storage device having an internal buffer.

[0002] External storage systems incorporating a so-called remote copy function have already been in practical use, in which data is mirrored and stored in disk subsystems of a main center and a remote center.

[0003] Such prior art has various issues to be solved because the remote copy function is realized by involving host computers.

"Synchronous Type and Asynchronous Type"

[0004] The remote copy function is mainly classified into two types, a synchronous type and an asynchronous type.

[0005] The synchronous type executes the following process sequence. When a disk subsystem is instructed by a host computer (upper hierarchical apparatus) of a main center to renew (write) data and if the disk subsystem is assigned the remote copy function, a renewal process completion notice is issued to the host computer of the main center only after the instructed data renewal (write) is completed for a corresponding disk subsystem in a remote center. A time delay (transmission time and the like) is generated in accordance with a geographical distance between the main center and remote center and the performance of a data transmission line therebetween. If the transmission time of the synchronous type is taken into consideration, several tens Km is a practical limit of a distance to a remote site.

[0006] In the synchronous type, the data contents in disk subsystems in the main and remote centers are always consistent from a macro viewpoint. Therefore, even if the function of the main center is lost by accidents or the like, the data contents immediately before the accidents are perfectly retained in the disk subsystems of the remote center and the process can be resumed quickly at the remote center. The term "always consistent from the macro viewpoint" means that during the execution of the synchronous type function, although the data contents may be different in terms of a process time (· sec, msec) of magnetic disk devices

and electronic circuits, the data contents are always the same at the time of data renewal completion. This is because the renewal process at the main center cannot be completed unless the renewal data is completely reflected upon the remote center. Therefore, in some cases, particularly if a distance between the main and remote centers is long and the data transmission line is congested, the access performance to a disk subsystem in the main center is considerably degraded.

[0007] In contrast, the asynchronous type executes the following process sequence. When a disk subsystem is instructed by a host computer of a main center to renew (write) data and even if this data is to be remotely copied, a renewal process completion notice is issued to the host computer of the main center immediately after the data renewal process for the disk subsystem in the main center is completed, to thereafter execute the data renewal (reflection) of the disk subsystem in the remote center, asynchronously with the data renewal in the main center. Since the data renewal is completed in a process time required by the main center, there is no transmission delay time or the like to be caused by storing the data in the remote center.

[0008] In the asynchronous type, the data contents in a disk subsystem of the remote center are not always consistent with those in the main center. Therefore, if the function of the main center is lost by accidents or the like, the data still not reflected upon the remote center is lost. However, an access performance of a disk subsystem in the main center can be maintained at the level when the remote copy function is not executed.

[0009] In order to back up data so as not to be lost by natural disasters such as earthquakes, it is necessary to set the distance between the main and remote centers to about 100 km to several tens km. Although it is possible to use a high speed communication line, for example, of a 100 Mbit/sec to 300 Mbit/sec class for the remote copy function, an expensive line subscription fee is incurred upon a customer of the disk subsystem, and this approach is not economically suitable.

"Order Integrity"

[0010] There is another problem different from the above-described issue of the data transmission time. Namely, if the remote center backs up the data of a plurality of disk subsystems of the main center, there occurs an issue (order integrity) that disk subsystems are required to be in one-to-one correspondence. In asynchronous remote copy, it is inevitable that reflection of renewal data in the remote center is delayed from the time when an actual renewal process is executed in the main center. However, the order of renewal in the remote center is required to be the same as that in the main center.

[0011] A database or the like is generally constituted of a main body of the database and various log and control information directly associated with the

main body. When data is renewed, not only the database main body but also the log and control information is renewed to maintain the system integrity. Therefore, if the renewal order is not kept, the integrity of information regarding the renewal order is also lost, and at the worst the whole of the database may be destructed.

"Involvement of Host Computer"

[0012] In the asynchronous remote copy under general environments where the main and remote centers have a plurality of disk subsystems, when the host computer instructs the disk subsystem to renew data, it is common that the host computer adds renewal order information such as a time stamp to the data to make the corresponding disk subsystem in the remote center execute a renewal data reflection process in accordance with the added information.

[0013] According to the remote copy function disclosed, for example, in the publication of JP-A-6-290125 (U.S. Patent No. 5,446,871), generation and supply of renewal order information and a renewal data reflection process based upon this information are realized through the cooperation between the operating system of a host computer in a main center and its disk subsystems and the operating system of a host computer in a remote center and its disk subsystems.

SUMMARY OF THE INVENTION

[0014] This prior art can realize the asynchronous remote copy function while ensuring the renewal order between main and remote centers. With this prior art, however, both the upper level software and a disk subsystem are required to have the mechanism for realizing the remote copy function, and also they are required to operate in cooperation. Since new custom software is required to be incorporated, a user is necessary to perform works such as software incorporation, setting and check, and modification of system designs to be caused by an increased CPU load. Incorporation of this conventional function is, therefore, associated with some obstacles such as a predetermined work period for such preparation and a cost therefor.

[0015] If the asynchronous remote copy function is executed when the capacity of the communication line to a remote center is not sufficient, renewal data not reflected upon the remote center increases.

[0016] It is an object of the present invention to realize an asynchronous type remote copy function capable of ensuring a renewal order and data integrity and facilitating its incorporation with less performance degradation of a main center, by using only the function of a disk subsystem without incorporating new software.

[0017] It is another object of the present invention to realize a remote copy function without incurring an expensive line subscription fee upon a customer of a disk subsystem, by applying an asynchronous type

remote copy function to the disk subsystem capable of storing a large amount of data.

[0018] Each of a main center and a remote center is provided with a disk subsystem serving as a gateway (hereinafter called a gateway subsystem) which is connected to a data transmission line. All disk subsystems in both the centers to which a remote copy is executed, are connected to the corresponding gateway subsystem of each center.

[0019] A volume of the disk subsystem in the main center to be remotely copied and a desired volume of the gateway subsystem in the main center are coupled by a synchronous type remote function to mirror data. If a system process time delay or the like can be neglected between the volume of the disk subsystem in the main center to be remotely copied and the volume of the gateway subsystem in the main center, data consistency can be retained.

[0020] Data is mirrored between volumes of the gateway subsystems of the main and remote centers through asynchronous remote copy. In this case, the gateway subsystem of the main center sends renewal data to the gateway subsystem of the remote center in accordance with the renewal order of volumes of disk subsystems of the main center, whereas the gateway subsystem of the remote center reflects the renewal data upon corresponding volumes of the remote center in accordance with the reception order of the renewal data.

[0021] Data is mirrored between the volume of the gateway subsystem of the remote center and the volume of each disk subsystem through synchronous remote copy. Data same as from a macro viewpoint is always stored in the volume of the gateway subsystem of the remote center and in the volume of the disk subsystem to be remotely copied.

[0022] The gateway subsystem stores data in the volume to be remotely copied, in a buffer memory of the gateway subsystem. Since the gateway subsystem has the buffer memory, generally an area for storing data in the gateway subsystem is not necessarily required. However, if there is an available area in the gateway subsystem, this area can be utilized for data transfer via the transmission line, depending upon the capacity of the transmission line.

[0023] With the above-described configuration, it is possible to mirror data between a plurality of disk subsystems of the main center and a plurality of disk subsystems of the remote center, by using the functions of the disk subsystems, while the data renewal order is retained. Reflection of renewal data upon the remote center can be performed asynchronously with the data renewal process at each disk subsystem of the main center. It is therefore possible to provide a disaster resistant back-up system of high performance and easy incorporation. Depending upon the communication capacity of the transmission line, the storage area of the subsystem can be utilized so that a burden of a line sub-

scription fee of a customer can be reduced.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024]

Fig. 1 is a diagram showing the overall structure of a remote copy system according to an embodiment of the invention.

Fig. 2 is a flow chart illustrating the details of an operation of the remote copy system.

Fig. 3 is a flow chart illustrating the details of an operation of the remote copy system to follow the flow chart of Fig. 2.

Fig. 4 is a flow chart illustrating the operation of a remote copy system having a gateway subsystem provided with a buffer area.

Fig. 5 is a diagram showing the internal structure of a gateway subsystem.

DESCRIPTION OF THE EMBODIMENTS

[0025] An embodiment of the invention applied to a general computer system will be described with reference to the accompanying drawings.

[0026] Fig. 1 shows an example of the structure of a system embodying this invention and allowing data to be mirrored between arbitrary two data centers among a plurality of data centers each provided with a general computer system.

[0027] One or a plurality of disk subsystems in a main center and one or a plurality of disk subsystems in a remote center are interconnected via gateway subsystems without involving the host computers, to realize a remote copy system for mirroring data between both the centers.

[0028] In the main center 12 shown in Fig. 1, a central processing unit (host computer) 1 is connected via interface cables 2 to disk subsystems 3-1, 3-2, ..., 3-n. The disk subsystems 3-1, 3-2, ..., 3-n store data to be referred to, or renewed, by the host computer 1. The gateway subsystem 5 is connected via interface cables 4 to the disk subsystems 3-1 to 3-n.

[0029] The gateway subsystem 7 is provided in the remote center 13 and connected via an interface cable 6 to the gateway subsystem 5 of the main center 12. The interface cable 6 is connectable to a general communication line. In this embodiment, therefore, the interface cable 6 is intended to include such a function.

[0030] When the host computer 1 issues a data write request to a disk subsystem 3-1 or the like, the disk subsystem 3-1 or the like writes the data in its buffer memory. Synchronously with this timing the disk subsystem 3-1 or the like issues a data write request to the gateway subsystem 5.

[0031] Upon reception of this write request, the gateway subsystem 5 writes the data in its buffer memory. Asynchronously with the data write in the buffer

memory of the gateway subsystem 5, the gateway subsystem 5 issues a data write request to the gateway subsystem 7 at the remote site. It is essential to use only one gateway subsystem 5 irrespective of how many disk subsystems 3-1 to 3-n are used.

[0032] The gateway subsystem 7 stores data supplied from the gateway subsystem 5 in its buffer memory, in the order of data write requests. It is essential to use only one gateway subsystem 7.

[0033] Disk subsystems 9-1, 9-2, ..., 9-n are connected via interface cables 8 to the gateway subsystem 7. When a data write request is issued from the main center 12 to the gateway subsystem 7, synchronously with this timing the gateway subsystem 7 writes the data therein and in the disk subsystem 9-1.

[0034] Remote copy is therefore executed by sequentially issuing a write request to a subsystem and then to the next subsystem. When a data write request is issued from the host computer 1 to one or a plurality of disk subsystems 3-1 to 3-n, the same data is loaded in one or a plurality of disk subsystems 9-1 to 9-n of the remote center 13. Arrows shown in Fig. 1 indicate the flow of the data instructed to be written by the host computer 1.

[0035] In the remote center 13, the host computer 11 is connected via interface cables 10 to the disk subsystems 9-1 to 9-n, and is a central processing unit which executes data reference and renewal relative to the disk subsystem 9-1 or the like. When the host computer 1 of the main center 12 cannot provide its intrinsic functions because of disasters, system failures or the like, the host computer 11 can operate as an alternative to the host computer 1. In addition, the host computer 11 can execute a process different from that of the host computer 1 of the main center 12 by using data stored in the disk subsystem 9-1 or the like, independently from the operation of the host computer 1. However, if the host computer 11 does not execute a process by using the disk subsystem 9-1, this host computer 11 is unnecessary.

[0036] The outline of the data mirroring method and operation according to the embodiment of this invention will be described with reference to Figs. 2 and 3.

[0037] A volume, a data set and a disk subsystem which store data to be mirrored are preselected by an administrator. The relation between the volume, data set and disk subsystem in which the data is stored and a volume, a data set and a disk subsystem which store a copy of the data, is preset to the disk subsystems by the administrator from the host computers.

[0038] In other words, a write destination of data to be copied is determined sequentially between respective disk subsystems described above. For example, when data is written in a volume of the disk subsystem 3-1, it is set in such a manner that in which volume of the gateway subsystem 5 the data is written from the disk subsystem 3-1, then in which volume of the gateway subsystem 7 the data is written from the gateway

subsystem 5, lastly in which volume of the disk subsystem 9-1 the data is written from the gateway subsystem 7. Such settings are provided for each disk subsystem including the gateway disk subsystem.

[0039] For such settings, the serial number and volume of each disk subsystem are used. For example, a volume A of the disk subsystem 3-1 is set to a volume B of the gateway subsystem 5, the volume B of the gateway subsystem 5 is set to a volume C of the gateway subsystem 7, and the volume C of the gateway subsystem 7 is set to a volume D of the disk subsystem 9-1. In this manner, the data written in the volume A of the disk subsystem 3-1 is copied to the volume D of the disk subsystem 9-1. This setting is conducted for all volumes.

[0040] Such preselect and preset may be effected by using a console or a service processor without using the host computer, if the disk subsystem can be connected to or provided with its console or service processor. The flow chart shown in Fig. 2 illustrates the operation assuming that the host computer is used for such preselect and preset.

[0041] For such preset, a specific address indicating the volume or disk subsystem may be designated, or the volume or disk subsystem in an arbitrary address range may be selected by using a control program in the disk subsystem. Path setting and pair setting are used as an example of initial setting (Fig. 2, 201).

[0042] The further description will be given with reference to the accompanying drawings.

[0043] As the host computer 1 (Fig. 1) issues a data write request (hereinafter called a write command) to the disk subsystem 3-1, 3-2, ..., 3-n (211) (Fig. 2, 202), the disk subsystem 3-1, 3-2, ..., 3-n executes a data load process for loading the write data therein in response to the write command, and also issues a write command for the data to the gateway subsystem 5 (212) (203). The write command is a command for transferring an instruction of data write and the write data itself.

[0044] Upon reception of the write command, the gateway subsystem 5 executes a process corresponding to the write command (204). After the gateway subsystem completes a data load process for loading the data in its buffer memory, it notifies a process completion to the disk subsystem 3-1, 3-2, ..., 3-n (211). A write command number is assigned to each write command in the order of the process completion (205), and at the timing determined basing upon the processing capability of the gateway subsystem 5, the write command assigned the write command number is issued to the gateway subsystem 7 (213) in the order of the write command number (206).

[0045] Under the conditions that the disk subsystem 3-1, 3-2, ..., 3-n completes a process for the write command issued from the host computer 1, i.e., completes a data load process for loading the data therein, and receives a write process completion notice from the

gateway subsystem 5 (212) (221), the disk subsystem supplies a write command process completion notice to the host computer 1 (222).

[0046] The gateway subsystem 7 (213) confirms, from the write command numbers assigned to respective write commands issued from the gateway subsystem 5 (212), whether the write commands have been received in the order of the assigned write command numbers. Thereafter, the gateway subsystem executes the processes corresponding to the write commands, i.e., executes a data load process (301) for loading the data in its buffer memory. Thereafter, a write command corresponding to the loaded data is issued to the corresponding disk subsystem 9 (311) (302). Upon reception of the write command issued from the gateway subsystem 7, the disk subsystem 9 (311) executes a process corresponding to the write command, i.e., executes a data load process for loading the data therein (303).

[0047] After the disk subsystem 9-1; 9-2, ..., 9-n (311) completes the process corresponding to the write command, i.e., completes the data load process for loading the data in its buffer memory, it supplies a process completion notice to the gateway subsystem 7 (321). Under the conditions that the gateway subsystem 7 (213) completes the data load process for loading the data therein and receives the write process completion notice from the disk subsystem 9-1, 9-2, ..., 9-n, the gateway subsystem 7 supplies a process completion notice for the write command to the gateway subsystem 5 (322).

[0048] According to the present invention, data written by the host computer 1 is mirrored between the disk subsystem 3-1, 3-2, ..., 3-n and the gateway subsystem 5 and is maintained consistent from a macro viewpoint. At this time, the gateway subsystem 5 adds information (serial number) to the data in order to hold the renewal order.

[0049] Data is mirrored between the gateway subsystems 5 and 7 through asynchronous remote copy while the renewal order is ensured. Synchronously with the data renewal by the gateway subsystem 7, the disk subsystem 9-1, 9-2, ..., 9-n renews the data. These operations are all realized by only the functions of the disk subsystem including a disk subsystem having the gateway function so that any load is not applied to the processing performance of the host computer.

[0050] The operation of the remote computer system which uses a buffer area of each gateway subsystem when the communication capacity of the transmission line is not sufficient, will be described with reference to Fig. 4. In Fig. 4, blocks having identical reference numerals to those shown in Figs. 2 and 3 have been already described above. In this system, a buffer area for temporarily storing write data is provided at each gateway subsystem in order to prevent an overflow of a buffer memory for a general transmission line. Data stored in the buffer area of the subsystem is sent from the main center 12 to the remote center 13 via the trans-

mission line, and to the gateway subsystem via the buffer area on the side of the remote center 12. Although time consistency for mirroring is degraded, the asynchronous type remote copy function can be realized without using a high capacity communication line.

[0051] Fig. 5 shows the structure of the gateway subsystem 5. The structure of the gateway subsystem 7 is the same as the gateway subsystem 5.

[0052] The gateway subsystem 5 has: an interface control unit 11 for data (including information) transfer to and from the disk subsystem 3-1 or the like and the gateway subsystem 7; a data buffer 12 for temporarily storing the data; a magnetic disk drive 13 as a storage medium for storing the data; a control memory for storing remote copy status information (as to which volume of which disk subsystem is written to which volume of the gateway subsystem 5, as to which volume of the gateway subsystem 5 is written in which volume of the gateway subsystem, and the like); a microprocessor 14 for controlling transfer of these data; a service processor panel 15 allowing a user to set how the remote copy is executed; and a disk array subsystem control unit 17 for controlling these components. In this example, although the data buffer 12 is provided at the gateway subsystem 12, this data buffer 12 is not necessary if a cache memory capable of performing a similar function to the data buffer 12 is provided, because the cache memory can function as the data buffer 12. In this specification, therefore, the data buffer 12 is intended to be inclusive of such a cache memory. Also in this example, although the control memory 16 is provided at the gateway subsystem 5, this control memory 16 is not necessary if a remote copy control information storage unit capable of performing a similar function to the control memory 16 is provided, because the remote copy information storage unit can function as the control memory 16. In this specification, therefore, the control memory 16 is intended to be inclusive of such a remote copy information storage unit.

[0053] As described so far, according to the present invention, an asynchronous type remote copy system can be realized which can ensure the data renewal order and data integrity by using the functions of disk subsystems without incorporating new software and which is easy to be incorporated and free from degradation of the process performance of the main center.

[0054] A storage area of the subsystem can be used depending upon the communication capacity of the transmission line so that a burden of a line subscription fee of a customer can be reduced.

Claims

1. A system having a plurality of subsystems (3, 9) connected to an upper hierarchical apparatus (1, 11), comprising:
a gateway subsystem (5, 7) having a buffer

memory for storing data supplied from said plurality of subsystems and an interface for connection to another system having other subsystems for writing the data therein.

2. A system according to claim 1, wherein said gateway subsystem stores data supplied from said plurality of subsystems and writes data into the other system, asynchronously.
3. A system according to claim 1, wherein said gateway subsystem adds order information of data to be transferred to the other system.
4. An integrated system comprising:
the other system (13) having a second gateway subsystem for receiving a write request sent from said gateway subsystem and a plurality of subsystems connected to the second gateway subsystem; and
the system (12) recited in claim 1.
5. An integrated system according to claim 4, wherein the other system is located at a remote site.
6. An integrated system according to claim 4, wherein the other system is connected by a communication line.
7. An integrated system according to claim 1 or 6, wherein the subsystem is a disk subsystem.
8. A data copy method comprising the steps of:
receiving a write request for data from an upper hierarchical apparatus (1) at a subsystem (3) of a first subsystem group;
writing the data in a first gateway subsystem (5);
writing the data written in the first gateway subsystem in a second gateway subsystem (7);
and
writing the data in a subsystem (9) of a second subsystem group from the second gateway subsystem.
9. A data copy method according to claim 8, wherein writing data from the upper hierarchical apparatus into the subsystem and writing data from the subsystem into the first gateway subsystem are executed synchronously.
10. A data copy method according to claim 8, wherein writing data from the subsystem into the first gateway subsystem and writing data from the first gateway subsystem into the subsystem of the second subsystem group are executed asynchronously.

11. A data copy method according to claim 10, further comprising a step of adding a number to data in the order of data write into the first gateway subsystem.

5

10

15

20

25

30

35

40

45

50

55

7

FIG.1

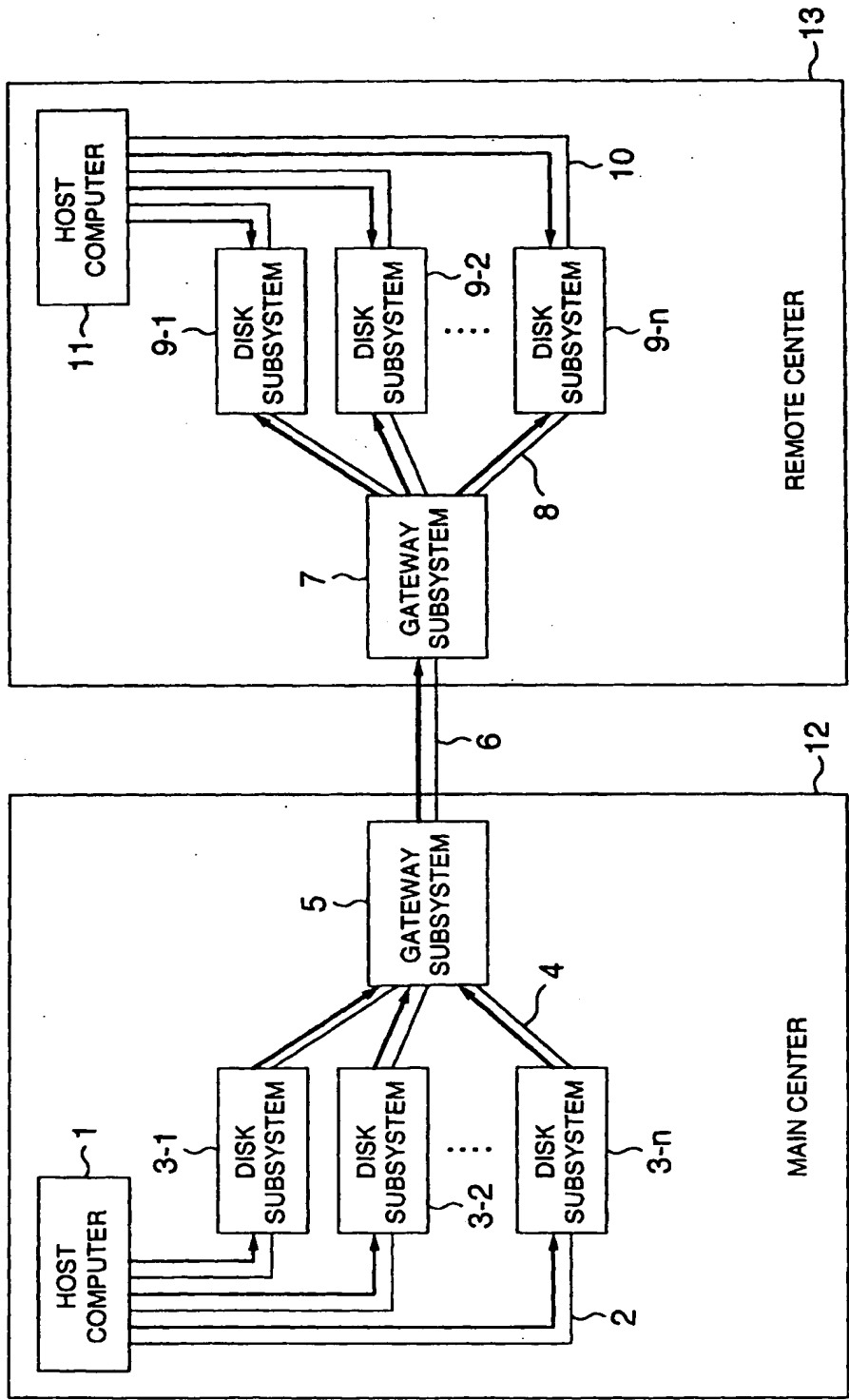


FIG.3

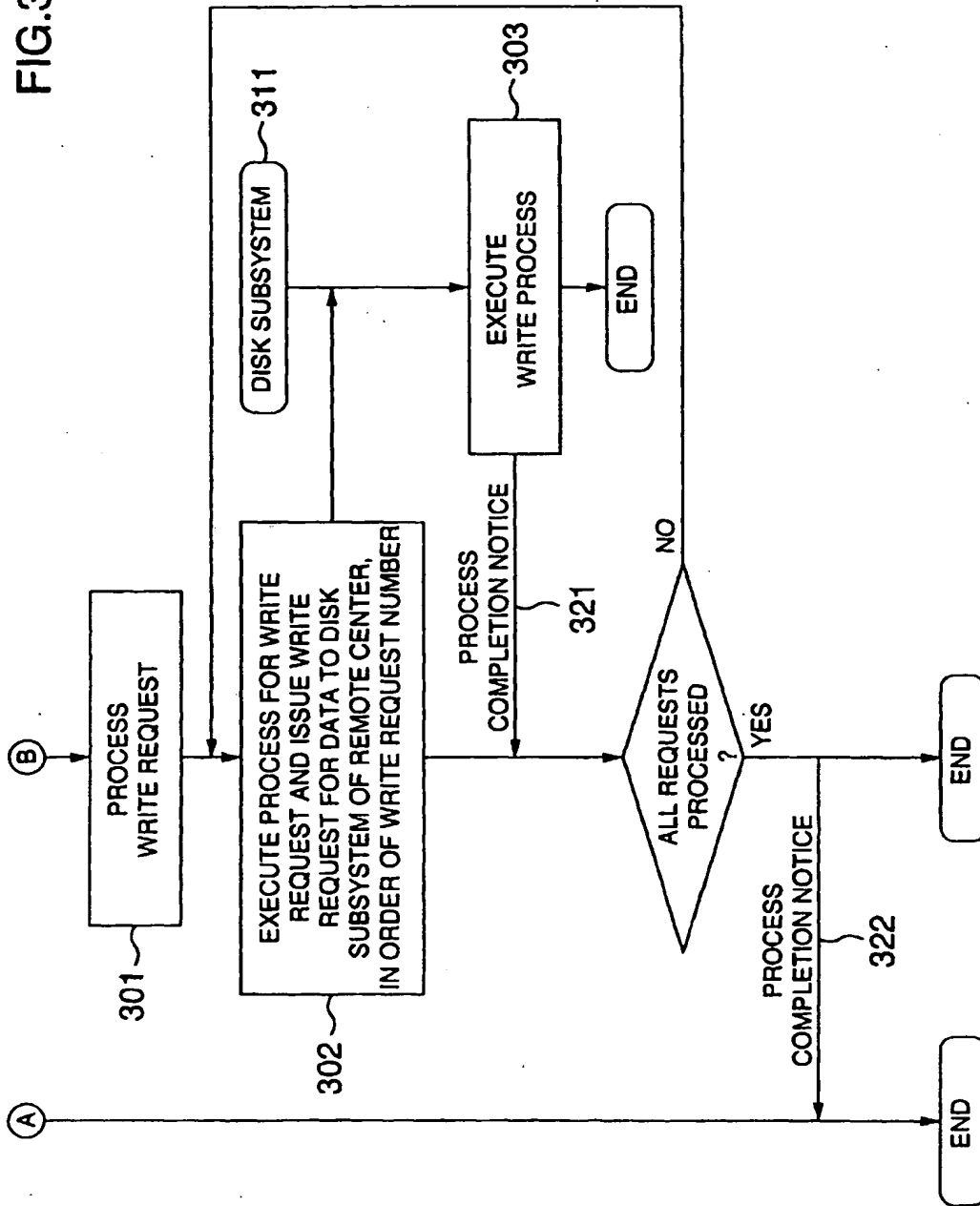


FIG. 4

